PRESENTATION

AI Factories Scaling AI-ready Infrastructure

1:35 PM - 2:00 PM

PRESENTER

Dave McCarthy

VP, Subdomain Lead, Cloud and Infrastructure Services, **IDC**

PRESENTER

Dave Pearson

VP, Subdomain Lead, Core Infrastructure, **IDC**

“

AI Factories are AI-ready infrastructure environments composed of accelerated compute, high-performance AI storage, intelligent networking fabrics, and AI-specific systems software designed to provide a standardized, scalable, and repeatable platform for training, fine-tuning, and deploying AI workloads across cloud, core, and edge.

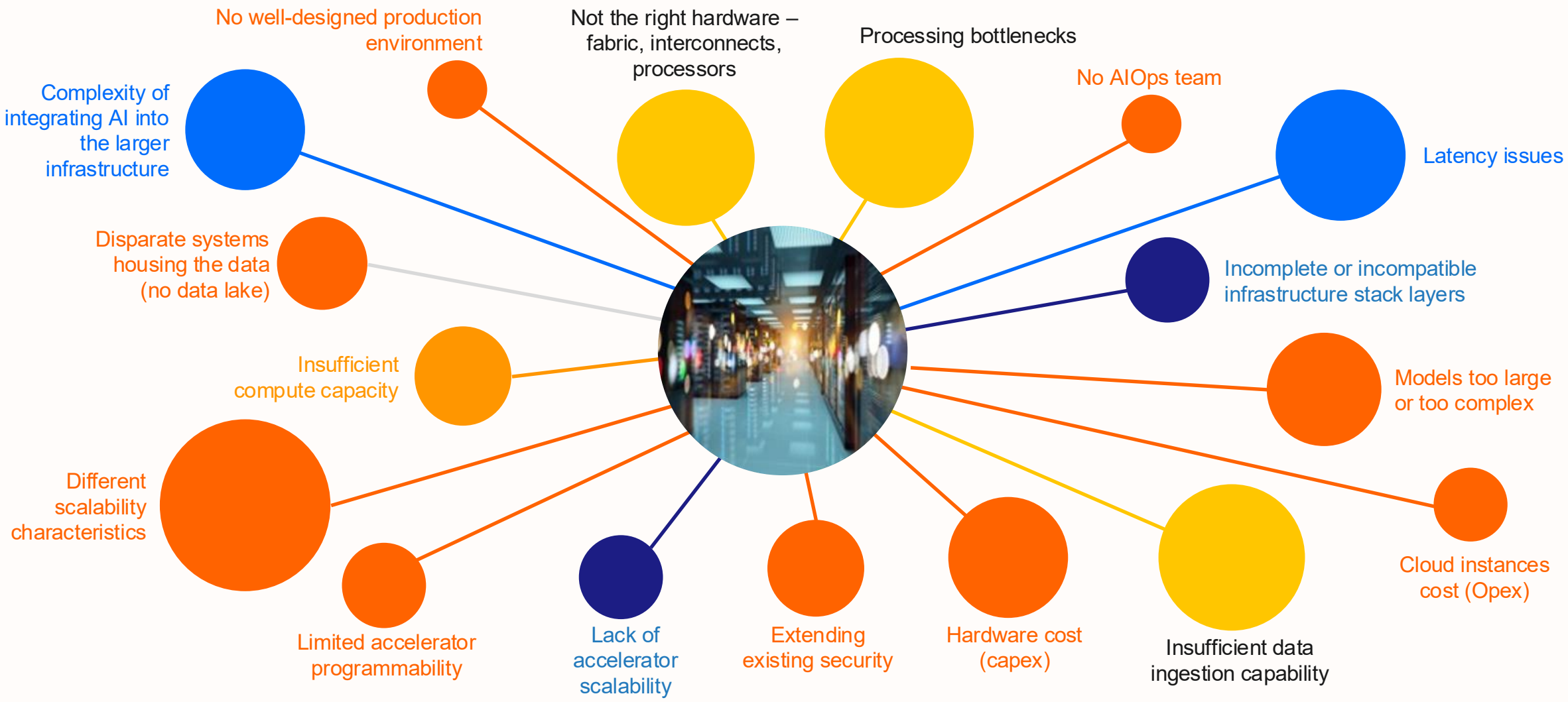
IDC’s definition of AI Factories

“By 2027, the massive computational and data demands of AI will compel 80% of organizations to modernize legacy cloud environments by shifting to new platforms specifically designed for AI workloads.

By 2028, 70% of the G2000 will modernize their data storage infrastructure and improve data logistics to feed quality, curated data to AI models on storage systems that can optimize GPU clusters.”

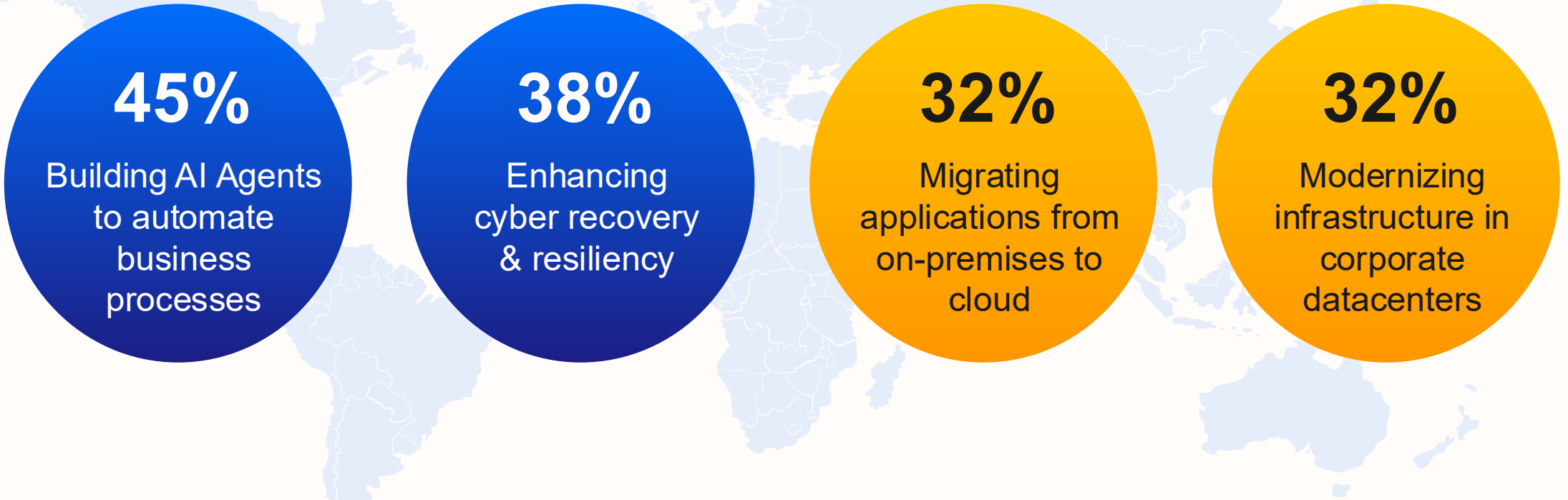
FutureScape Predictions, 2026

IT buyers tend to underestimate the role played by infrastructure



Migration to cloud and modernization of datacenters are of equal interest to enterprise customers

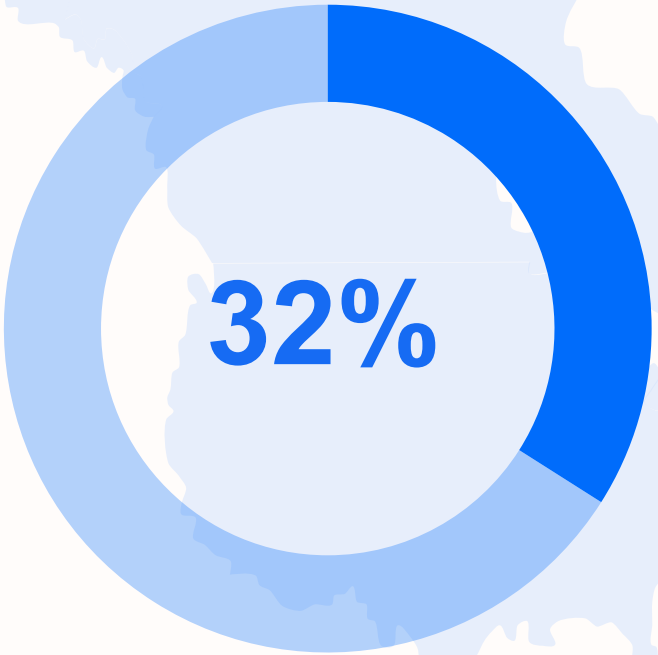
Top Areas for Significant Budget Increases in 2026, Regardless of Overall Budget Plans



Source: Future Enterprise Resiliency & Spending Survey Wave 9 IDC, December 2025, N=1,007

Regardless of overall IT spending plans, in which areas do you expect to most significantly increase spending in 2026?

Migrating applications from on-premises to cloud (IaaS or SaaS)



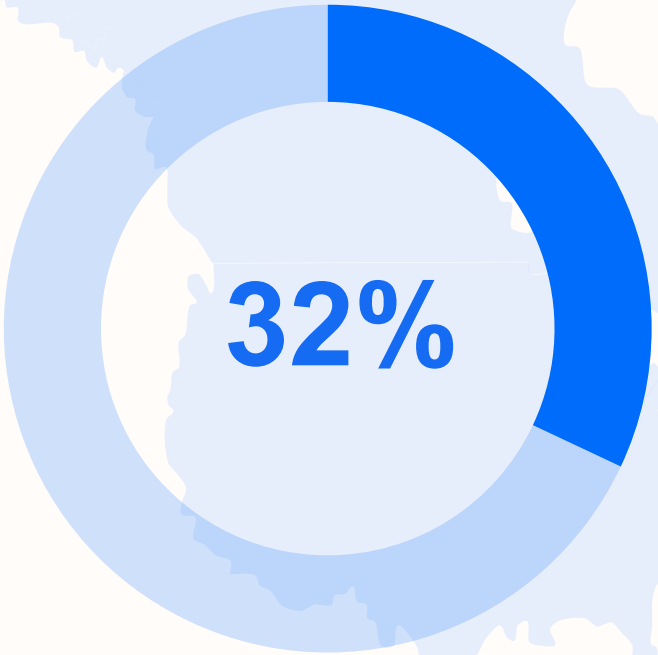
Most important outcomes you want to achieve from your migration to cloud

- 45% Ensure workloads are easier to AI-enable
- 37% Improving the cyber-resiliency of workloads
- 37% Ensure workloads can easily use high performance infrastructure option
- 31% Make it easier to augment core applications with agentic extensions
- 25% Ensure workloads can easily use low-cost infrastructure options

Future Enterprise Resiliency 9 Spending Survey Wave 9 IDC, December 2025, N=364

Regardless of overall IT spending plans, in which areas do you expect to most significantly increase spending in 2026?

Modernizing infrastructure in corporate datacenters

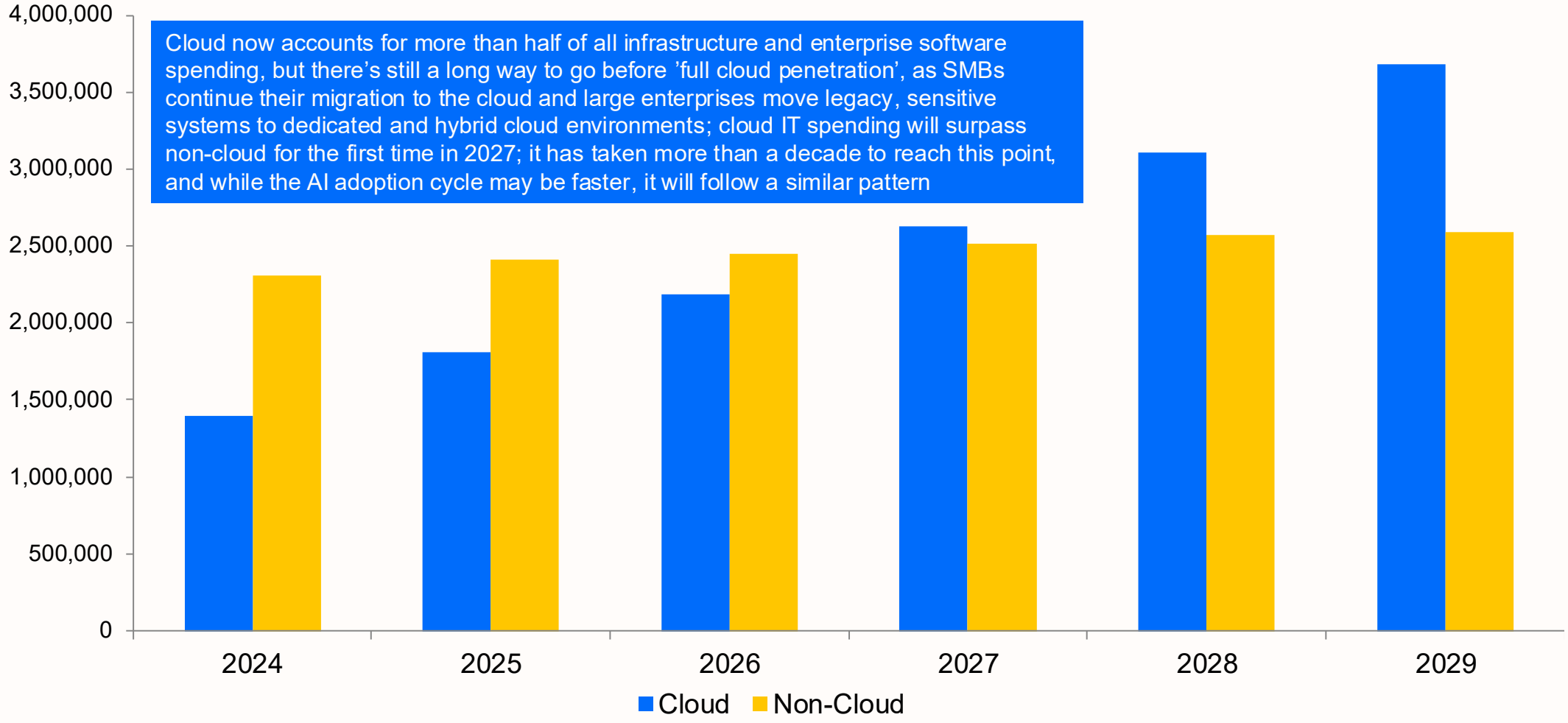


Future Enterprise Resiliency 9 Spending Survey Wave 9 IDC, December 2025, N=364

Top priorities for modernizing infrastructure SW (e.g., virtualization, containers, OS) in 2026

- 60% Investing in infra-automation and AI-enabled observability solutions
- 50% Accelerating efforts related to mainframe modernization
- 40% Increasing use of open-source infrastructure software
- 30% Using bare metal to support AI and other higher performance workloads
- 19% Shifting from virtual machines to container-based applications

Cloud migration continues to accelerate but datacenters are not shrinking



Source: Worldwide Black Book, December 2025 (growth in constant currency); Cloud includes all spending on public cloud services and software (IaaS, SaaS, PaaS); cloud infrastructure (public and dedicated cloud); and cloud-related IT services (public and dedicated cloud); non-cloud includes everything else

What is AI sovereignty?

AI sovereignty is a dimension of an AI strategy that describes the ability to have **free choice and control** over the design, development, deployment, accessibility, operation, maintenance and governance of AI systems and applications, and the technologies that those systems and applications depend on.

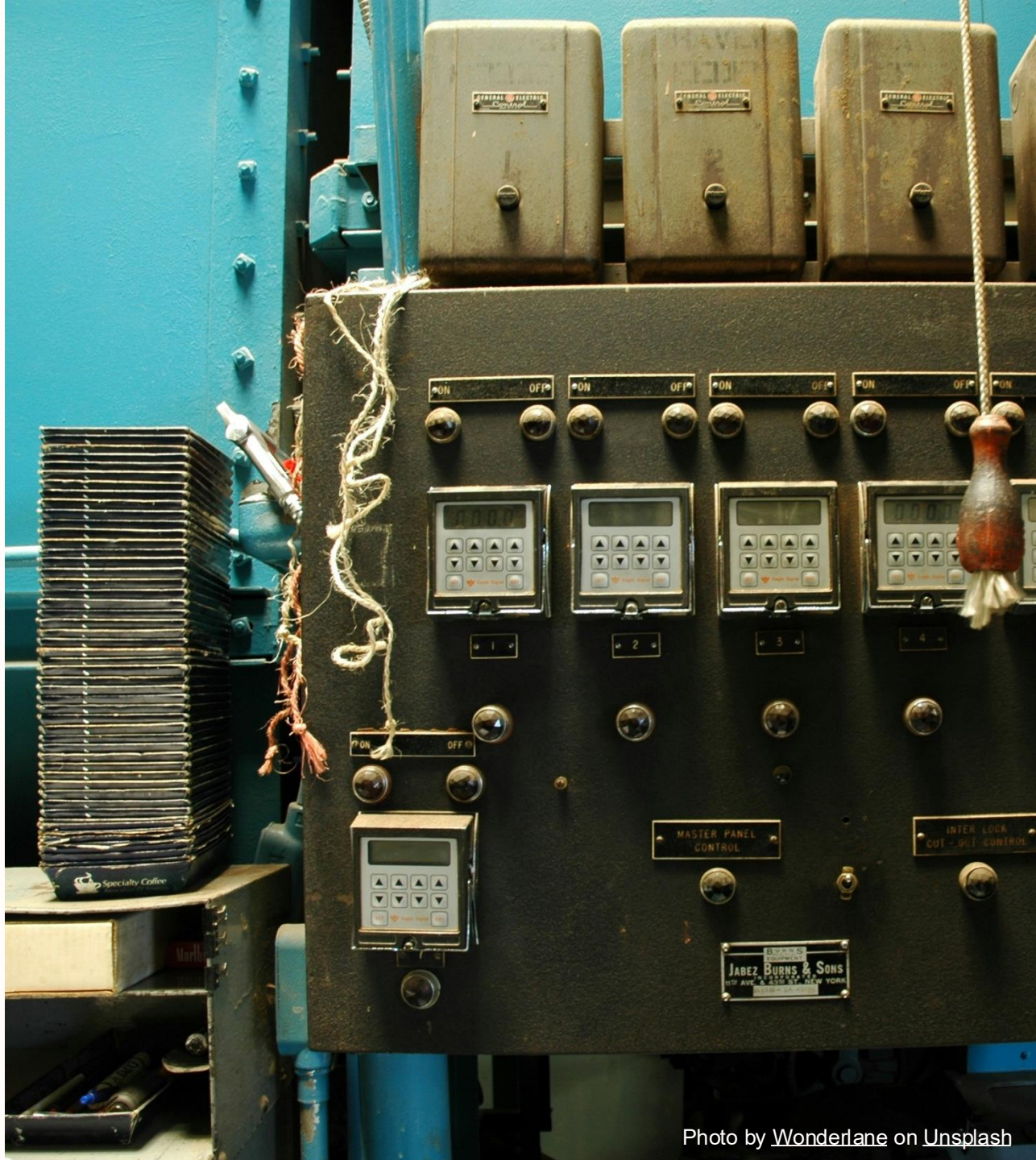


Photo by Wonderlane on Unsplash

AI sovereignty-related enterprise concerns are common — and should be taken seriously.

86%

of organizations want freedom to choose platforms other than public cloud for AI inference at scale.

29%

are concerned about data or IP loss due to improper AI implementation/use.

58%

plan to use open models to power generative AI (GenAI) use cases.

27%

cite AI investment decisions being impacted by regulatory uncertainty

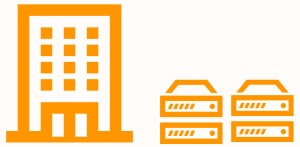
IDC Predicts:

By 2027, 75% of non-U.S. G2000 enterprises will prioritize pursuit of AI sovereignty, using a blend of non-public hosting, open technologies, and regional partners to support mission-critical AI uses.

Sources: IDC's *AI Technology Trends Survey*, June 2025; *Future Enterprise Resiliency & Spending Survey Wave 7*, Sept 2025; IDC EMEA *AI-Ready Infrastructure Survey*, June 2024; *IDC FutureScape: Worldwide AI & Automation 2026 Predictions*

Hybrid infrastructure approaches bond together multiple deployment strategies

Self-built and/or managed infrastructure



Deployed in traditional data center, co-lo or dedicated cloud infrastructure

- Data confidentiality
- Private AI
- Customized analytic and dev stacks

Computing platforms and systems

NVIDIA, intel, WEKA.IO, cisco, DELL, ddn

Storage systems

XILINX, AMD, FUJITSU, IBM, ORACLE, Red Hat

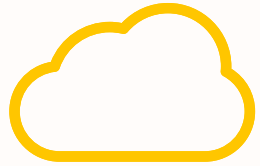
Storage and computing **Infrastructure software**

Hewlett Packard Enterprise, vmware, Lenovo, SUPERMICRO

and others

Interoperability and workload mobility

Infrastructure as a Service



Deployed as dedicated or public cloud infrastructure as a service

- Consumption-based pricing and support
- Standard APIs
- Built-in advanced analytics and dev services

Public and dedicated cloud IaaS (compute and storage)

IBM, ORACLE, Hewlett Packard Enterprise

PaaS

aws, Microsoft Azure, DELL, Lenovo

SaaS

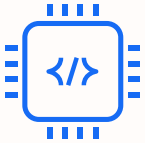
Google Cloud

and others

Essential guidance

Define Your Value Proposition

- Demonstrate where to find quick wins and the greatest ROI for AI
- Help isolate common requirements across use cases



Right-size infrastructure offering

- Optimize performance and scale for cost-effective AI and agent deployments
- Rebalance workloads across public cloud and datacenter

Enable sovereignty and governance

- Help guide integrated strategy across units and functions
- Account for the multiple layers of sovereignty: data, operational, and technical

Highlight your hybrid strengths

- Consistent operations from data center to edge to cloud
- Capex and opex options for both dedicated and shared resources

IDC

DIRECTIONS