Analyst Brief, sponsored by:
Penguin Solutions

Building Sovereign AI Infrastructure: A C-Suite Guide to Control, Compliance, and Competitiveness

Written by:

Rahiel Nasir
Research Director

Published:
September 2026

AI workloads create new points of failure, turning digital sovereignty into a boardroom infrastructure decision. This brief guides C-suite leaders on control, compliance, and competitiveness in sovereign AI infrastructure.

Introduction

When dependency becomes an operational risk

Most enterprises know exactly what happens when a datacenter fails or a cloud region goes dark. Failover engages, the incident is logged, and service resumes. Far fewer can say what happens when a foundation model API is withdrawn, a GPU allocation is cancelled, or an AI vendor loses its export license. These exposures are not variants of an outage. The continuity processes already in place will not mitigate them because those processes were built for infrastructure that fails rather than for access that is revoked. This gap is the real story behind digital sovereignty in 2026.

At a Glance

- About 81% already use, or plan to use within 12 months, sovereign cloud to build or run AI workloads.
- About 51% cited data privacy and security as the leading driver of sovereign cloud adoption, ahead of pure regulatory compliance.
- About 63% said dependency on foreign AI providers is a major or moderate geopolitical concern.

Digital sovereignty has moved from a data residency compliance exercise to a boardroom infrastructure decision. IDC's Worldwide Digital Sovereignty Survey finds that 89% of organizations globally already use sovereign cloud for at least some workloads or plan to within 12 months, and that 81% of them use or plan to use it specifically to build or run AI. What has changed is not the enthusiasm but the scope. AI training and inference workloads concentrate risk around compute, data, and model dependencies in ways earlier cloud workloads did not, exposing organizations to failure points such as model access cutoffs, export restrictions, and sudden changes to provider terms. Many operational playbooks were never built to handle these exposures.

This reframes sovereignty as four related disciplines. Data and model sovereignty comes first, covering control over the data used to train and fine-tune models and over the models themselves so that both remain governed by local law and organizational policy. Technical sovereignty covers the infrastructure and compute an organization controls directly. Operational sovereignty asks who runs that environment and who is accountable when something goes wrong. Skills sovereignty asks whether an organization can secure and retain the skills needed to operate a sovereign environment, which is a different question from operating without outside help.

Few enterprises will run sovereign AI infrastructure alone, and a growing number of national sovereign AI programs require that technical operations be carried out by local citizens or residents, which makes access to qualified, credentialed people a sovereignty requirement in its own right rather than a gap to be closed internally. Together, these four disciplines form a deliberate infrastructure strategy that meets sovereignty requirements while preserving the scale and performance AI initiatives demand and turns sovereignty into a competitiveness and continuity issue for the C-suite rather than a compliance footnote for legal and IT.

Definitions

Organizations use the following terms inconsistently. IDC's definition draws on IDC's Worldwide Digital Sovereignty Taxonomy:

- Digital sovereignty refers to an organization's or nation's ability to control its own digital destiny, spanning data, technology, operations, and skills.
- Data sovereignty refers to data that is subject to, and protected by, the laws of the jurisdiction in which it is collected or stored.
- Cloud sovereignty (or sovereign cloud) refers to cloud infrastructure and services architected to meet data, operational, and jurisdictional control requirements.
- AI sovereignty (or sovereign AI) refers to control over the data, models, and infrastructure used to build and run AI so outcomes remain governed by local rules and organizational policy rather than an external party.
- Operational and skills sovereignty refers to the extent to which an organization can run and support its own sovereign environment, including its ability to secure personnel who meet local residency or citizenship requirements where these apply.

Benefits of sovereignty

Handled deliberately, sovereign AI infrastructure delivers more than regulatory cover. It builds regulatory and compliance confidence, reducing exposure to extraterritorial data requests and aligning with local and sector rules. It builds operational resilience, enabling AI operations to continue through geopolitical disruptions, export restrictions, or a sudden change in a provider's terms, precisely the scenarios most organizations have not yet stress-tested. Proactive, automated approaches that keep GPU-based clusters available and job-ready reduce the operational burden that sovereignty can otherwise impose, without compromising performance.

It protects competitive intellectual property (IP) as well. About 98% of organizations now treat ownership of their training data as critical or important, recognizing that proprietary data and fine-tuned models are a source of advantage, not only an operational input. Finally, sovereignty commitments are becoming a trust signal. In regulated industries and cross-border relationships, showing where data lives, who can access it, and which laws govern it increasingly matters as much to customers and partners as the underlying AI capability itself.

Considerations

None of this is free. High cost (35%) remains the most cited challenge in implementing sovereign cloud, followed by concerns over network security and data in transit (34%), high complexity (33%), managing IT and vendor risk (32%), and integration with existing multicloud environments (31%).

A continuum, not a switch

Sovereignty is not a binary choice or a switch an organization flips. It sits on a continuum. IDC describes it as a layered stack spanning applications, platforms, and infrastructure, crossed with assurance levels running from basic IT controls through to governance and customer-facing guarantees. C-suite leaders can use this continuum as a practical good, better, best framework for evaluating control requirements. Good might mean contractual data residency commitments. Better adds jurisdictional operational control. Best extends to dedicated or air-gapped infrastructure. Each step trades cost and deployment time for a firmer guarantee, and the right tier depends on whether the workload is subject to regulatory compliance or uses highly sensitive data. Organizations should choose the most appropriate IT venues for their workloads rather than treat a single sovereign environment as an organization-wide standard.

The trade-off is quantifiable. About two-thirds of organizations are prepared to pay a premium of up to 20% on top of total public cloud spending for sovereign cloud, with the 6–10% band as the most commonly cited. That tolerance is expected to tighten rather than widen: 12 months out, 61% expect to pay no more than 15%. More instructive is where the money goes. Specialist skills and talent (39%), new data governance tooling (39%), new cybersecurity solutions (38%), and sovereign or local AI platforms and model hosting (37%) all rank ahead of local infrastructure and platforms (25%) as areas needing extra budget. The cost of sovereignty is therefore mostly operational rather than capital, which is precisely why the skills and operational dimensions deserve the same board-level attention as the infrastructure decision itself.

Ownership versus access

On infrastructure choice, organizations lean toward dedicated, owned, and operated environments. Among organizations already using or planning sovereign cloud, 37% currently run sovereign private or dedicated cloud and a further 41% intend to within 12 months, whereas fewer than 5% rule it out altogether. Respondents were asked about each platform type separately and could select more than one, so these shares describe the mix of environments in use rather than competing slices of a single total. On this basis, sovereign private and dedicated cloud is on course to become the most widely deployed sovereign environment during 2027, and infrastructure planning should assume that shift rather than react to it.

Air-gapped environments, used by 26% today, with a further 33% planning adoption within 12 months, sit at the most restrictive end of the same spectrum. Every platform type covered by this question is sovereign by design, so an air-gapped deployment is not an alternative to sovereignty but its strictest expression. It therefore sits within the sovereign cloud footprint already counted here rather than alongside it. These platform shares are drawn from all organizations that use or plan to use sovereign cloud, a wider group than the 81% applying it specifically to AI workloads.

Separate IDC research on IT procurement finds that leasing on-premises GPU servers has become an important strategy for keeping AI costs within budget while meeting strict security requirements, reinforcing this lean toward ownership. One caveat is that IDC's findings vary across sectors and geographies. IDC's research in the government and education sectors

found sovereign and private AI preferences to be more common among organizations still exploring AI than among those operating at scale, while IDC's research in Japan found that early adopters plan to increase their use of private AI infrastructure as they mature. Both findings reflect specific sectors and markets rather than global patterns. Whether organizations' preference for sovereign AI infrastructure decreases as they mature in their AI adoption remains an open question. Buyers should also watch for sovereign washing, in which sovereignty is used as a marketing label without real technical or operational substance behind it, and weigh the trade-offs between best-in-class global AI capabilities and tighter control that sovereign infrastructure provides.

Trends

Four developments will shape this market over the next 12 to 18 months.

Geopolitical tension is broadening organizations' reassessment from cloud dependency to technology provider dependency generally, and this concern is not evenly distributed. IDC's 2026 *Worldwide Digital Sovereignty Survey* shows that 71% of organizations in Asia/Pacific and 76% in Japan called dependency on foreign AI providers a major or moderate concern, compared with 58% in North America. This is a reminder that any infrastructure strategy needs a genuinely global view rather than a European or North American default.

Sovereignty considerations are also moving up the AI stack, from infrastructure to models and platforms, prompting some organizations to explore distributed, federated architectures in which models travel to the sovereign zone to train rather than restricted data moving to a central model. IDC's own research on AI-fueled business strategies describes this fragmentation of the AI stack as a structural feature of 2026, not a temporary phase. IDC continues to track how it affects integration cost and deployment time.

Meanwhile, national and regional investments in in-country compute capacity continue to expand, reshaping where large AI workloads can realistically run, most visibly in the Middle East, Southeast Asia, and parts of Europe.

Finally, evolving regulation, from data protection regimes to emerging AI-specific rules, is raising the bar for what counts as a credible sovereign AI control set. Sovereign controls are becoming a stated purchase criterion for SaaS and platform selection, not merely an infrastructure layer consideration.

Conclusion

Sovereignty has become an infrastructure strategy question that sits with the C-suite, not a compliance footnote handed to legal and IT. The practical starting point is a map. Which infrastructure, data, and operational dependencies would actually break if a single AI vendor or jurisdiction became unavailable tomorrow and who in the organization is accountable for keeping the business running if it did.

From there, leaders should treat operational and skills sovereignty as seriously as data residency when selecting infrastructure and partners and use a tiered good, better, best view of sovereignty to match the level of control to the value and sensitivity of each workload,

rather than applying a single standard everywhere. Organizations that translate sovereignty requirements into a proactive, tiered infrastructure strategy protect both compliance and competitiveness. Those that treat it as a box to tick are the ones most likely to discover, at the worst possible moment, exactly where their new points of failure were.

Practical considerations to guide sovereign AI deployment

- Assess sovereignty across all four dimensions: data and models, technical infrastructure, operations, and skills.
- Treat sovereignty as an infrastructure and business resilience strategy, not solely a legal or compliance exercise.
- Use a tiered good, better, best framework to align sovereignty requirements with workload sensitivity and risk.
- Evaluate deployment and operating models on the level of control, accountability, and operational independence each one delivers.
- Score prospective partners against the criteria buyers cite most often: country-level security and cloud certifications (36%), sovereign control of network infrastructure and connectivity (31%), ownership of in-country datacenters (31%), robust IT risk management (30%), and dedicated AI sovereignty capabilities (28%).

Unless otherwise stated, all data in this document is drawn from IDC's *Worldwide Digital Sovereignty Survey*, June 2026, with 633 respondents.

TABLE 1

Top challenges in implementing sovereign cloud ranked by share of organizations citing the challenge

Challenge	Share of organizations (%)
High costs	35
Network security and data-in-transit concerns	34
High complexity	33
Managing IT and vendor risk	32
Integration with existing multicloud environments	31

Note: n = 633

Source: IDC's *Worldwide Digital Sovereignty Survey*, June 2026

About the IDC analyst(s)

Rahiel Nasir, Research Director, IDC

Rahiel is a research director on IDC's cloud and datacenters, enterprise infrastructure practice and leads IDC's Worldwide Digital Sovereignty research program. He has covered technology markets for all of his career, including a prior role as a datacenter infrastructure and services analyst. He also worked as a journalist, covering enterprise and consumer technologies.

More about Rahiel Nasir

Message from the sponsor

About Penguin Solutions

Penguin Solutions (Nasdaq: PENG) is an AI factory platform company that designs, builds, and manages next-generation datacenters for enterprises, sovereign AI initiatives, and neocloud providers.

With deep design expertise at the intersection of datacenter AI infrastructure and memory solutions, our full-stack AI factory platform combines differentiated infrastructure software, advanced memory, compute systems, end-to-end services, and industry-leading partner technologies to help customers accelerate deployment, optimize token economics, and maximize the return on their AI investments.

Learn more at PenguinSolutions.com.

IDC Custom Solutions

IDC Custom Solutions produced this publication. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis that IDC independently conducted and published, unless specific vendor sponsorship is noted. IDC Custom Solutions makes IDC content available in a wide range of formats for distribution by various companies.

This IDC material is licensed for external use, and in no way does the use or publication of IDC research indicate IDC's endorsement of the sponsor's or licensee's products or strategies.

IDC Research, Inc.

One Beacon Street, Suite 33100, Boston, MA 02108, USA

T +1 508 872 8200

blogs.idc.com | LinkedIn @IDC | www.idc.com

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. With more than 1,300 analysts worldwide, IDC offers global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries. IDC's analysis and insight help IT professionals, business executives, and the investment community make fact-based technology decisions and achieve their key business objectives.

©2026 IDC. Reproduction is forbidden unless authorized. All rights reserved. CCPA