AI Innovation: 5 Key Pillars to Enable Sovereign AI

Learn about the 5 pillars of sovereign AI: (1) AI capabilities, (2) data residency, (3) data privacy, (4) legal controls, and (5) security and resiliency

November, 2024, Version 1.0

Table of contents

Introduction

Governments and private organizations are adding AI to their digital strategy to accelerate innovation, optimize operations, and stay competitive in a rapidly evolving digital landscape. However, adopting AI technologies raises concerns around security, privacy, and trustworthiness, particularly in cloud environments. Consequently, AI adoption may be closely tied to establishing a strong digital sovereignty stance, allowing companies to maintain control over their data. The term "sovereign AI" has emerged as a descriptor of these needs.

In this article, we present the key tenets that Oracle believes cloud providers should offer to help address sovereign AI requirements for both public and private entities. These five pillars help ensure that all organizations can select the desired digital sovereignty guardrails for their use case, without hindering their ability to drive AI-powered innovation.

The idea of digital sovereignty isn't binary— it isn't something you either have or don't have. Instead, it's more like a spectrum. It evolves over time and needs to be constantly refined to adapt to changes in the digital world. Every AI environment is unique, and should be analyzed based on its unique characteristics. We'll outline sovereign AI requirements across five pillars: (1) AI capabilities, (2) data residency, (3) data privacy, (4) legal controls, and (5) security and resiliency. Across these pillars, we'll highlight some of the areas where Oracle Cloud Infrastructure (OCI) is uniquely positioned to help customers in their journey toward sovereign AI.

1. AI capabilities

AI portfolio management

AI technologies are available to casual business users, coders and developers, and everyone in between. Cloud technology providers should offer a breadth of AI capabilities to meet the needs of all users. While underlying AI infrastructure is crucial for performance and the user experience, relatively few end users will require direct access to the hardware itself.

Large language models (LLMs) and foundational models (FMs) form the backbone of generative AI (GenAI). These models are developing continuously and expanding their use cases for customers. Alongside these developments, innovations in AI computing technologies are crucial to achieving the cost efficiencies required to sustain this rapid progress. New GPU processors are being introduced multiple times a year. Thousands of AI models exist for a plethora of use cases. Several dozen LLMs have emerged as industry leaders, and they all regularly get updates and release new versions.

It's becoming an increasingly fragmented landscape, where the complexities of selecting the right hardware and software are compounded by the need to continually update, upskill, and reevaluate your AI portfolio. Partnering with a hyperscaler as both a strategic advisor and a cloud vendor can help provide clarity in your selection of infrastructure and AI models. A good hyperscaler will enable reduced costs by providing the latest GPUs as needed while also helping reduce the need to continually retrain your team on constantly evolving infrastructure and AI models. In addition, leveraging AI infrastructure and services in the cloud lets businesses experiment with cutting-edge AI technologies, driving innovation without needing to build infrastructure in-house. This lets organizations focus on developing AI solutions rather than managing the underlying hardware.

Below are some of the key AI portfolio management capabilities that OCI offers.

AI infrastructure

Training large GenAI workloads, such as LLMs, often requires substantial infrastructure resources, which can be expensive to run and should be optimally configured for efficiency. OCI is the only major cloud provider to offer bare metal instances with state-of-the-art GPUs for high performance that are free of virtualization overhead. Checkpointing, the process of periodically saving the internal state of an AI model during training, is a critical capability that can save time and money and reduce downstream model errors. For checkpointing during AI training, OCI's instances provide the most local storage per node (61.4 TB). For a balance of performance and price, OCI Compute Virtual Machines (VMs) with GPUs are also available and consistently less expensive than ones from AWS and Azure.

OCI can achieve terabits per second of throughput in a single file system with OCI File Storage and its new high performance mount target feature. Remote Direct Memory Access (RDMA) is a technology that lets one computer directly access another computer's memory without the involvement of either computer's operating system or compute resources—that's extremely useful for high-throughput, low-latency networking in massively parallel compute clusters. OCI's native RDMA clustering lets customers scale from three GPUs in a single-node edge appliance to the largest AI supercluster in the industry, with more than 131,000 GPUs.

AI services and SaaS applications

Organizations will require AI services and SaaS to access the capabilities that GPUs provide.

With OCI, customers can access a comprehensive suite of AI services, including cutting-edge generative AI and machine learning innovations. Additionally, Oracle Fusion Cloud SaaS applications come equipped with embedded AI functionalities, enabling seamless integration of intelligent features. Every layer of the stack leverages the unique capabilities of OCI AI infrastructure. As a result, customers using these services are positioned to benefit from enhanced performance, flexibility, and AI-driven solutions tailored to their needs.

A platform with complete cloud services

AI applications live in an interconnected ecosystem of applications; therefore, the ideal AI environment should offer a robust set of cloud services to complement infrastructure and AI model capabilities.

Oracle Cloud's Everything Everywhere commitment offers hyperscaler functionality from the edge to the core to the cloud. With Oracle's distributed cloud, customers can deploy OCI's more than 150 AI and cloud services at the edge, in their own data center, across clouds, or in the public cloud, and can help address a variety of data privacy, sovereign AI, and low latency requirements. In addition, Oracle Cloud is available in more locations than any other hyperscaler, with more than 80 regions live and 77 planned.

2. Data residency

Data residency is a key component of sovereign AI and refers to the common requirement that data must remain within the physical boundaries of a particular nation or jurisdiction. In the context of digital sovereignty, we refer to jurisdiction as the specific area or region where a legal authority, such as a court or government, has the power to enforce laws and make decisions. For example, the European Union consists of multiple countries under a common legal authority. Regarding sovereign AI, data residency requirements may mandate that data used to train, operate, and improve AI systems is stored, processed, and managed locally.

AI use cases requiring the use of sensitive information, such as personally identifiable information (PII), raise a fundamental question about data residency. When organizations use common LLMs, such as ChatGPT, Claude, LaMDA, PaLM, or Jasper AI, there's a risk of sensitive information leakage. This can occur during the training of the models. LLMs are trained on massive data sets, which can include sensitive information. This data remains within the model's architecture, even after training is complete and the data set is deleted. All AI-driven data and metadata processing, including inputs and outputs to LLMs, may then need to be securely contained within a single country or jurisdiction, depending on applicable requirements. Using AI with sensitive data across global cloud infrastructure could potentially result in noncompliance with data residency regulations.

In OCI, data and customer-generated metadata is restricted to a single Oracle Cloud region by default. Unlike other hyperscalers, the vast majority of OCI services are regional, meaning that there are very few back-end processes that require communications with other regions. Only critical security services, such as key management, tagging, and identity management are cross-regional, but other core services aren't. Using a zero trust security approach, by default, customers have access only to the selected home region and can intentionally enable the use of other OCI regions where data can be stored and processed via policy configuration. In addition, Oracle Cloud offers at least two regions in every jurisdiction where it operates in order to provide sovereign disaster recovery.

Physical separation from general-purpose regions

Oracle's distributed cloud portfolio offers realms across different regions and cloud types. A realm is a group of connected Oracle Cloud regions. Realms are completely separate from one another, meaning they don't share any information. A customer's environment exists in a single realm, enabling them to access all regions that belong to that realm—but not any other regions outside of it. Oracle Government Cloud, Oracle EU Sovereign Cloud, Oracle Cloud Infrastructure Dedicated Region, Oracle Alloy, and Oracle Cloud Isolated Region realms are all physically, logically, and cryptographically separated from the rest of Oracle's public cloud regions. Data residency becomes intrinsic to the chosen platform. This architecture helps prevent unauthorized data transfer outside of the customer's jurisdiction by design and offers a comprehensive set of safeguards that enhance data privacy and help you address legal compliance—features that are unique to OCI.

3. Data privacy

Data privacy helps ensure that sensitive data remains confidential and protected from unauthorized access. In the context of sovereign AI, maintaining privacy is essential to helping prevent foreign bad actors or malicious entities from accessing or misusing data.

The use of AI models and their outputs, specifically when used to aid decision-making by humans or within apps, brings a range of data privacy risks. As an example, if an unauthorized individual gains access to a company or government's proprietary data contained within their LLMs, they could potentially extract sensitive information. If an LLM is prompted with a question that requires it to generate text, there's a risk that it could inadvertently disclose sensitive data it was trained on. Even if the LLM doesn't explicitly mention sensitive information, it may be possible to infer that information based on the model's responses. For example, if an LLM is trained on a data set that includes classified documents, it may be able to generate text that's consistent with the content of those documents, even if it doesn't explicitly mention them.

Oracle manages fined-tuned models and helps ensure access limitations. The models leverage Oracle's security capabilities, regardless of the vendor that provides the base model. OCI also offers encryption at rest and in transit and includes a native integration with third-party key management systems, helping ensure that data involved in the AI-based lifecycle (training, inferencing, and outputs) can be encrypted with keys managed outside of the Oracle Cloud regions.

OCI is also the only hyperscaler that offers bare metal GPU compute. When an OCI bare metal instance is provisioned, it has no OCI control plane elements, and the entire instance is controlled by the customer. This adds an additional layer of isolation and privacy that no other provider offers.

Cleared cloud operations and support

Dedicated Orcale Cloud regions can be operated and supported by an independent team with cleared personnel to help you meet specific regulation requirements. As an example, European customers that may need local operations can choose Oracle EU Sovereign Cloud regions in Frankfurt, Germany, and Madrid, Spain, where access is granted exclusively to EU residents hired within a dedicated EU Sovereign Cloud legal entity for deployment, operations, and security.

Facility ownership

OCI can build a dedicated Oracle Cloud region specifically for a single customer—bringing a full region into the customer's data center, behind their firewalls to help enforce both residency and operational control. This deployment model is also available to organizations that want to partner with Oracle and become cloud service providers for their own customers with Oracle Alloy.

One customer per cloud region

As an elevated privacy safeguard with OCI Dedicated Region, only a single customer has exclusive access to a cloud region. Similarly, Oracle Alloy allows only a single partner to access the region, but that partner can host multiple customer cloud environments within it, retaining control on who can access the cloud environment.

Data isolation

When workloads require the highest level of protection to handle highly classified and regulated data, Oracle offers Oracle Cloud Isolated Region, which is completely disconnected from the internet and OCI, accessible only in an "air gap." Oracle Cloud Isolated Region is designed to meet the higher demands of global customers' mission-critical classified workloads and deliver the same set of more than 150 OCI cloud services in an on-premises environment.

4. Legal controls

Cloud computing represents a significant shift from traditional on-premises models. In traditional setups, organizations maintain full control over their hardware and software stack, managing infrastructure in-house. In contrast, the cloud involves utilizing resources from a third-party service provider, where certain components are controlled externally while others remain the organization's responsibility. This introduces a shared responsibility model, where security and privacy must be jointly managed by both the cloud customer and the service provider, ensuring collaboration across the technology stack for optimal protection and governance.

Many countries have strict data privacy laws that regulate how data can be classified, collected, processed, and stored. For example, in Europe the General Data Protection Regulation (GDPR) protects the privacy of individuals by helping ensure their personal data is handled in a secure and lawful manner. In addition, the European Union AI Act is designed to promote the adoption of trustworthy AI while ensuring certain high-level protections. Similarly, all around the world, regulatory frameworks are emerging that are designed to address digital sovereignty concerns. Noncompliance could result in significant legal, financial, and reputational consequences. With more than 80 compliance certifications—spanning GDPR, the Federal Risk and Authorization Management Program (FedRAMP), the National Institute of Standards and Technology (NIST), and the EU Cloud Code of Conduct—aimed at alignment with even the strictest regulatory requirements, Oracle Cloud enables customers to address the regulatory policy within their selected geography.

EU Sovereign Cloud legal entities

In Europe, Oracle EU Sovereign Cloud regions rely on bespoke, dedicated, EU-based sovereign legal entities that own the hardware and data center leases and provide the operations and support with EU-resident teams. Oversight for the EU-based legal entities is provided by a governance committee facilitating fidelity with current and future regulations. In addition, EU Sovereign Cloud is supported by specific agreements, such as the data processing agreement, the service description, and a dedicated addendum to the service pillar document. These contracts outline responsibilities for handling personal information, third-party subprocessors, confidentiality, and security standards. These contracts are designed to help address the requirement that customer content will not leave the selected EU Sovereign Cloud region(s) without a customer's authorization or instruction and to reduce the risk of unauthorized access by entities or individuals outside the EU Sovereign Cloud organizations.

5. Security and resiliency

Like all online environments, AI environments, including their training data and model parameters, can be vulnerable to cyberattacks, which could result in data breaches. These mission-critical environments may also require redundancy and resilient design against outages or regional disasters to avoid data and service losses and to enable continuity in case of catastrophic events.

The deployment of AI models, especially when used to assist decision-making, whether by humans or within applications, can introduce a spectrum of cybersecurity challenges. Customers want to safeguard data used for training and fine-tuning and to defend against adversarial attacks designed to manipulate AI predictions. The complexity of the AI stack, including hardware, software, and code transparency, can make cybersecurity a major challenge. Further complicating these issues: While governments and organizations alike are reaping the benefits of new AI capabilities, cybercriminals are using AI to enhance their illegal activities. The proliferation of deepfakes and increasingly sophisticated phishing threats emerge from generative AI misuse. These risks, increased by potential identity theft from AI-generated content, demand robust cybersecurity strategies to foster responsible AI usage and governance.

Customers can help protect their data with a set of integrated cybersecurity solutions in Oracle Cloud, leveraging the same technology that received Department of Defense Impact Level 6 certification, the standard for the highest level of classified data in the US. In addition, OCI provides advanced data governance tools (such as Oracle Cloud Guard, Oracle Data Safe, and OCI Audit) to help customers manage and audit their LLMs and FMs with all the related data during the entire lifecycle in the cloud environment.

Dual-region resiliency

OCI offers resiliency with more control over data residency, with two regions offered in many countries. Oracle EU Sovereign Cloud, Oracle Government Cloud, and OCI Dedicated Region all use the same multiregion design. This two-region (i.e. Frankfurt and Madrid for EU Sovereign Cloud) approach offers protections against regional disasters and general backup/failover support while also maintaining customer data within the borders of the jurisdiction, enabling customers to meet even the most stringent region-based residency compliance policies.

Distributed database

Oracle Globally Distributed Autonomous Database is a fully automated distributed cloud database that helps customers address their data residency requirements, provides survivability for business-critical applications, and delivers cloud-scale database performance. It's a single logical Oracle Database that's distributed as a set of Oracle Database shards. Each shard is an independent Oracle Database instance that hosts a subset of the logical database data. Multiple shards can run in one OCI availability domain to maximize performance, in multiple OCI availability domains or regions to provide the highest possible availability, and in OCI regions across different geographies to help meet data residency requirements. In other words, data can be distributed across multiple regions while also allowing sensitive data with residency or other compliance restrictions to stay only in the mandated jurisdiction.

In addition, OCI helps customers with automated recovery services, such as OCI Full Stack Disaster Recovery and Oracle Database Zero Data Loss Autonomous Recovery Service, to simplify the implementations of business continuity strategies.

Getting started: Implementing the five pillars of sovereign AI

AI is transformative, but it may need to be leveraged while maintaining a solid digital sovereignty stance, especially when it's deployed in the cloud. Innovative AI solutions can lead to increases in revenue, reduced costs, and enhanced operational efficiency for governments and enterprise organizations globally, but new technology isn't without risks. It's critical to consider residency, privacy, security, and legal controls when implementing AI solutions. Oracle's distributed cloud is uniquely positioned to help customers implement their AI strategy while helping them meet their digital sovereignty requirements with the industry's broadest set of sovereign capabilities.

Learn more about Oracle sovereign AI, or speak to an Oracle AI expert to implement the five pillars of sovereign AI and enable AI innovation for your organization.

Connect with us

Call +**1.800.ORACLE1** or visit **oracle.com**. Outside North America, find your local office at: **oracle.com/contact**.