ORACLE

7 Data Infrastructure Questions for AI Success

By addressing these key points, leaders can simplify AI adoption, maximize ROI, and power business innovation

Table of contents

Is your data ready for AI?

By Jeffrey Erickson
Senior Writer

Imagine the same large language models, or LLMs, that can explain quantum physics and summarize French novels gaining deep expertise on your company's unique operational data and knowledge base. Now retrieving, combining, and contextualizing that data moves from a task in an analyst's queue to a natural language conversation between a business user and an AI agent that can uncover insights, then take actions. IT resources are freed up and your organization is more data driven.

Sounds good, right? So it's no wonder that the world's most ambitious, talented, and well-funded tech companies—and their startup competitors—are racing to get the power of generative AI into your hands. Even amid this hype, it's hard to overstate the extent to which GenAI models can multiply the value of your data by fundamentally altering how people throughout your organization access and use it.

The foundation to all this is an infrastructure keyed to the needs of generative AI—one with the right mix of customized tools, foundational LLMs, techniques, and compute systems to provide fast responses and support use cases as varied as anomaly detection and object identification. Success often depends on having enough power to handle complex AI operations—think GPUs interconnected by a cluster network that's able to achieve ultra-high performance and microsecond latency. It also requires vector databases; a way to leverage AI agents for retrieval-augmented generation, or RAG, which combines your diverse enterprise knowledge bases with LLMs; and user-friendly AI agent interfaces.

> It's hard to overstate the extent to which GenAI models can multiply the value of your data.

What’s Agentic AI?

Agentic AI refers to artificial intelligence that’s capable of understanding and responding to information as well as actively pursuing objectives.

Key characteristics of agentic AI

Proactive behavior
The AI can initiate actions rather than simply reacting to external prompts.

Adaptation
The AI can learn from experience, accept feedback, and adjust to better achieve its goals.

Goal-oriented behavior
The AI has specific objectives it seeks to achieve and can map out steps to get there.

Autonomy
The AI can make decisions and take actions independently, within parameters.

The promise of GenAI is substantial. So are the infrastructure investments needed to get there

You’ve probably heard eye-popping stats on data center expansions to support AI. Much of that spending has been by hyperscale cloud providers looking to deliver stellar AI experiences, with some forward-looking organizations also prepping their data infrastructures to support what’s coming.

So what can you do to get ready now? Below are seven questions many IT architects are asking as they work to deliver on the promise of AI now and into the future.

1 How will we benefit from a range of GenAI use cases, including some no one's thought of yet?

A handful of AI projects—think hyperpersonalized marketing, really good customer service chatbots, productivity-boosting coding helpers—will often justify the cost of the supporting infrastructure. And the list of potential use cases is growing by the day, because frankly, if you can imagine it, you can probably do it. That's largely thanks to the advanced LLMs that are now embedded in enterprise applications, such as ERP, HCM, and SCM, and widely used in research-heavy areas such as robotics, healthcare, genomics, and aerospace engineering. In other words, nearly everywhere.

The key is to supplement those LLMs so your AI becomes the leading expert on your business. By providing access to historical, operational, financial, and documentation data libraries, you let GenAI serve as a force multiplier in many ways for many people throughout your enterprise.

What does that look like? It can mean executives "chatting" with data, drilling down and deriving new business intelligence on the spot. It can mean AI agents becoming partners for software development and R&D brainstorming sessions, even lightning-fast creators of prototypes and mockups. It can mean salespeople capably managing more leads because prospecting, communications, and process minutia get new levels of automation, courtesy of AI.

These are just a few examples. A brainstorm session where your product, finance, HR, and legal teams share their ideas will likely yield creative GenAI use cases and a solid starting point for a data infrastructure plan.

Want more ideas from peer companies? Check out a range of real-world use cases as varied as genetic analysis and sports broadcasting.

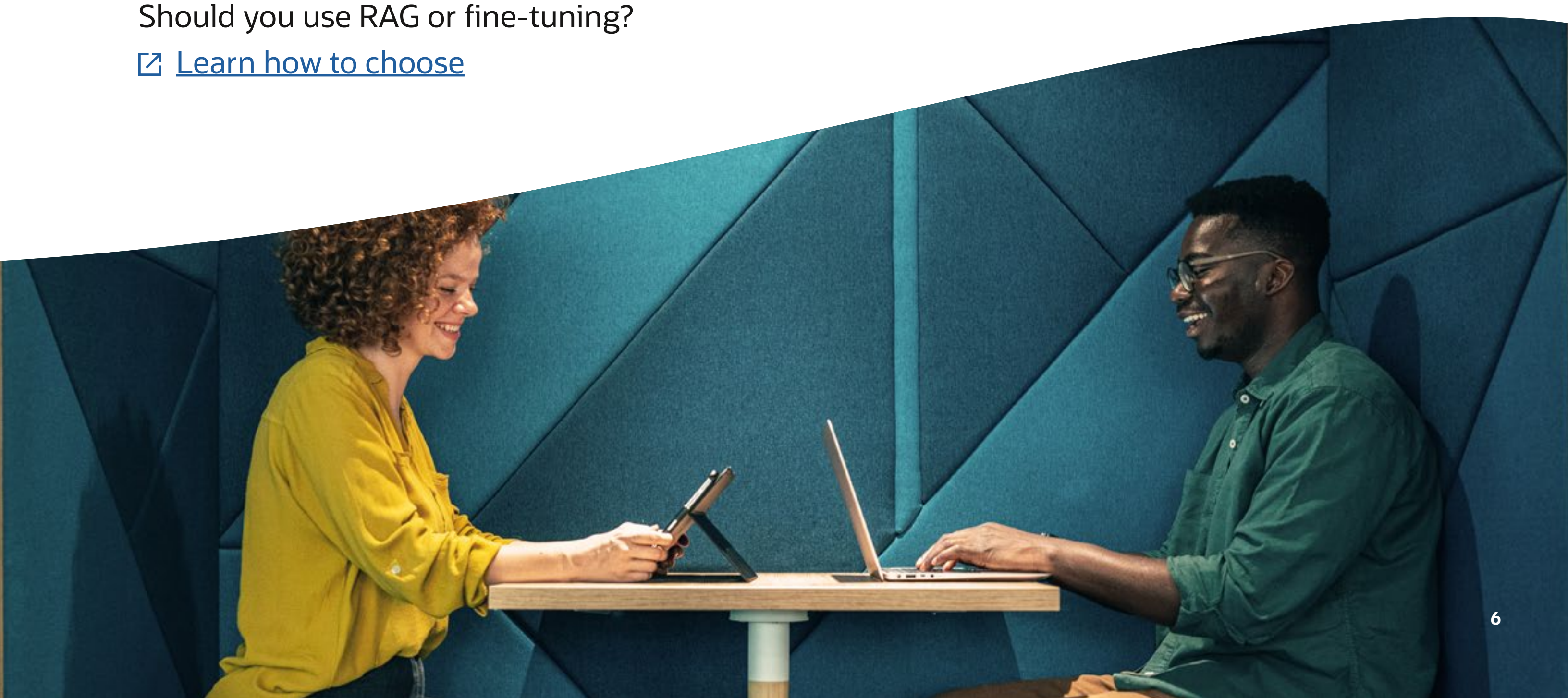
Fine-tuning and RAG: Two ways to help AI “get” your data

Both fine-tuning and RAG help generative AI models deliver responses that are more contextually relevant and tailored to your organization. Here’s how they differ.

Fine-tuning means taking a general-purpose model, such as Command from Cohere or Llama 3 from Meta, and giving it additional rounds of training on a smaller, domain-specific data set. Tuning helps the model perform better on specific tasks because it’s been adapted to the nuances and terminology of a domain, such as coding, finance, or healthcare. Downsides are cost and the need for a GPU-based infrastructure to support the process.

RAG is an architectural framework that helps general-purpose AI models deliver outputs useful to specific organizations by providing selected, relevant data to the LLM as it formulates answers to queries. The result is an AI system that combines the language fluency of an LLM with insights from your internal data to deliver targeted, more contextually appropriate responses. RAG, unlike AI model fine-tuning, works without modifying the underlying model. The LLM will also “forget” the data provided to it after queries are complete, alleviating a potential source of data leaks.

Should you use RAG or fine-tuning?
Learn how to choose



2 How will we make all of our unstructured and semi-structured data available for AI?

Processing unstructured and semi-structured data for use by AI is a multistep process that involves data collection and ingestion, storage, and processing—cleansing, feature extraction, normalization, segmenting, and possibly other steps. Only then are the images, documents, and audio and video files in your data lake prepared for fine-tuning, vectorization, and RAG.

Data for fine-tuning: Fine-tuning GenAI models for specific tasks can require showing it new discipline-specific data. For instance, if you're tuning a generative AI model to produce medical reports, you'll need a data set of high-quality, relevant medical reports to help the model learn the proper terminology and context.

Data for ongoing AI outputs: An AI-ready data management system needs to provide your LLMs with access to a wide range of data stores, such as JSON documents and relational and semi-structured data. It needs to efficiently apply vector embeddings to data, store the embeddings as vectors in a database, and query the database to enable efficient vector search and natural language processing. This can require ongoing data collection and an integration pipeline that also includes a RAG architecture to help improve the accuracy and relevance of your AI model outputs.

What does it look like when these systems are in place? A sales team, for example, could save the audio of every call and let AI analyze customer sentiment while creating summaries. No more indecipherable scribbling as salespeople try to engage and take notes at the same time. No forgotten details on a quote request for the customer's next order. It's the same attention to detail, but effortlessly more accurate and efficient.

One key to success: A unified, multimodal database that lets you enable all these processes without moving, resynchronizing, and resecuring data across specialized systems. Such a unified data model can speed up operations by orders of magnitude, allowing for searches across data types, including AI vectors, without IT needing to integrate disparate silos. This can lead to richer results as AI models consider subtle relationships between lots of seemingly unrelated but critical information.

3 How will we secure sensitive data for AI, preferably with a single governance and access control strategy?

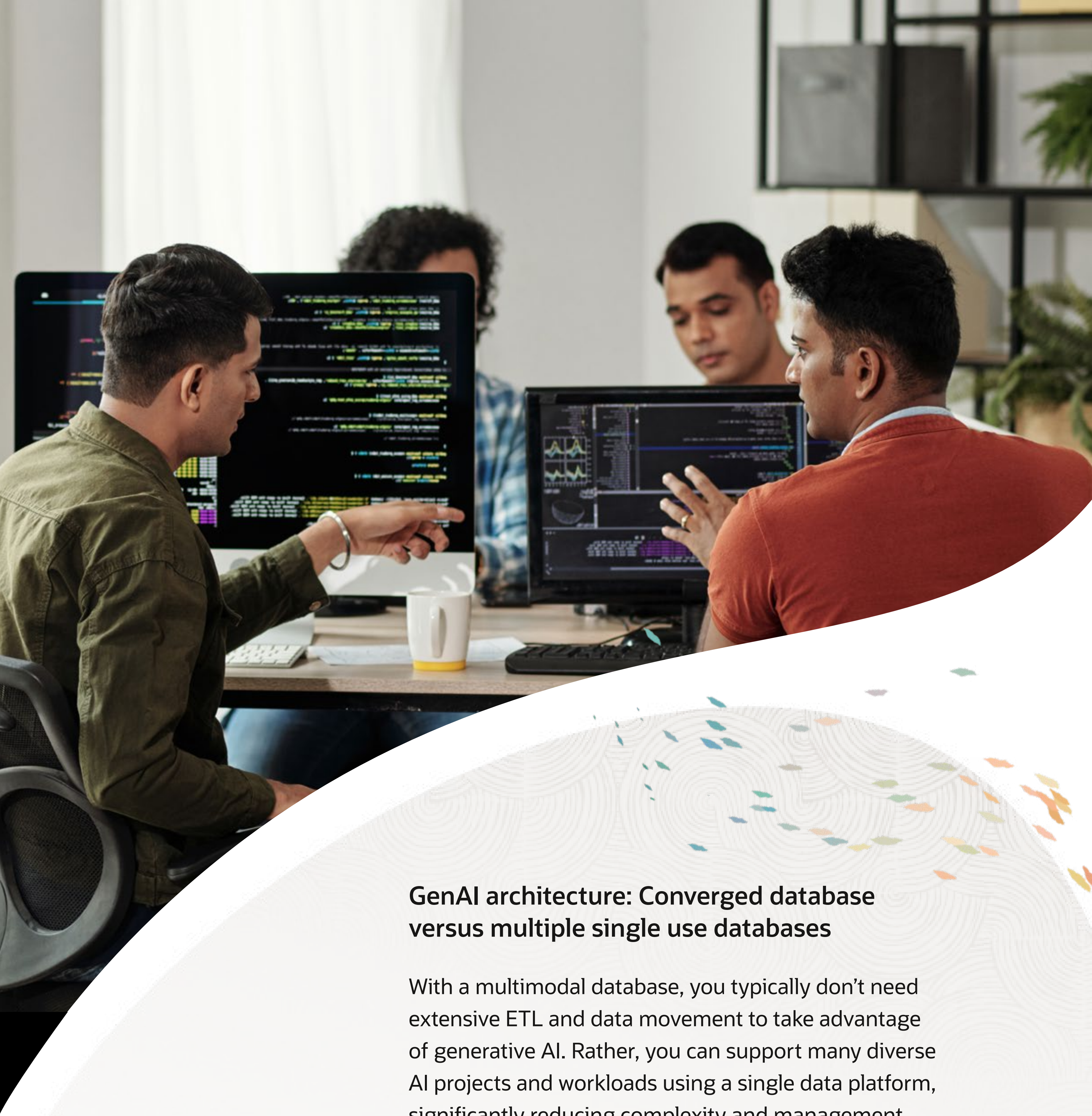
Maintaining the security and privacy of the data provided to your generative AI tools is an essential infrastructure design tenet, in both model fine-tuning and RAG and during daily use.

Depending on your needs, data for your generative AI platform can be anonymized, encrypted, or masked. Data destined for use in generating AI outputs also needs to be tightly controlled by considering the requester's role and employing multifactor identification processes. Some organizations choose to run inferencing on their own copies of AI models housed on dedicated infrastructure, or even in on-premises data centers, to further protect private data.

You can look to your technology providers and standards bodies for assistance. For example, Oracle is among the more than 280 organizations collaborating with the National Institute of Standards and Technology (NIST) in its Artificial Intelligence Safety Institute Consortium[1]. The AISIC's goal is to "develop science-based and empirically backed guidelines and standards for AI measurement and policy, laying the foundation for AI safety across the world." Members will work to develop guidelines, tools, methods, protocols, and best practices to help the community develop and deploy AI securely.

There are other, similar AI-centered efforts, too.

[1] *NIST Artificial Intelligence Safety Institute Consortium*



GenAI architecture: Converged database versus multiple single use databases

With a multimodal database, you typically don’t need extensive ETL and data movement to take advantage of generative AI. Rather, you can support many diverse AI projects and workloads using a single data platform, significantly reducing complexity and management overhead while minimizing the risk involved in moving data. A multimodal database doesn’t mean data must be in one format or model; it just means you can choose your data architecture for each type of application.

4 How will teams collaborate on our GenAI strategy while keeping standards in place?

In the right hands, an off-the-shelf model can become a customized AI powerhouse that only your organization could create. People want to take advantage of AI. Giving them the tools they need to do so makes it much more likely that GenAI workflows will become part of your company's strategy—and not one-off shadow IT projects.

One way to maximize your efforts is a center of excellence, or CoE, that provides transparency and promotes consistency across departments. A CoE consolidates best practices, tools, and techniques related to generative AI and leads the way in using them to solve business problems.

ORACLE
CHECKLIST
How to Build Your AI Center of Excellence
Do you have the right AI resources, skills, and strategies to stay ahead of the competition?

Learn how to build an AI CoE in your company

A core technology for this collaboration will be a data science platform that encourages collaboration for the technical steps of fine-tuning and deploying models. Your hyperscale cloud provider will have a generative AI platform where data scientists can work with and enhance powerful foundation models from Cohere, Meta, and other suppliers. These platforms give data scientists and IT teams a place to store, exchange, and potentially reuse models, data sets, and data labels across services.

Hyperscalers also offer model catalogs and provide fine-tuned LLMs with the data and compute infrastructure to run them. Now business units can piggyback on previous AI successes in other parts of the organization.

With both a CoE and a leading data science platform, you can foster a culture of continuous learning and improvement. Providers including Oracle offer fully managed services, such as Oracle Cloud Infrastructure (OCI) Generative AI, for seamlessly integrating LLMs into a wide range of use cases. You can also create custom models by fine-tuning base models with your own data set.

5 How can we combine the strengths of our cloud providers to maximize data availability for AI?

Maybe your organization is exploring Google Gemini or Microsoft's Copilot. Or you could decide to host open source models in AWS or Cohere's foundation models in OCI. Now, new agreements between Oracle and the other hyperscalers open a wide world of possibilities for companies whose corporate data governance strategies are built around Oracle Database technology, on-premises or in the cloud. You can embrace any of these models with low latency and management overhead, thanks to innovative multicloud relationships between Oracle and Azure, Google Cloud, and AWS that let you use Oracle Database services not just in OCI, but inside each provider's data centers.

Success with those not-yet-conceived generative AI use cases may be more likely when you can use exactly the AI models you want, running where you want, without compromising on easy and secure access to your data. Teams can leverage the best services for specific tasks while maintaining security and resiliency and reducing costs.

6 How will we procure, manage, and afford the systems we need for fine-tuning and inferencing?

Fine-tuning and deploying generative AI is a compute-intensive undertaking—each interaction can involve complex calculations on massive data sets where an LLM containing billions of parameters draws on specialized computing systems to power its manipulation and analysis of information. The more intricate the model, the more computation it needs to process data and deliver relevant outputs. You'll need a plan to keep compute costs down for both model fine-tuning and inferencing while delivering high-quality results.

Rather than incurring the costs and developing the expertise to build a system in-house, many organizations select popular data science platforms, such as those available from major cloud providers, to make data sets clean and relevant to a task.

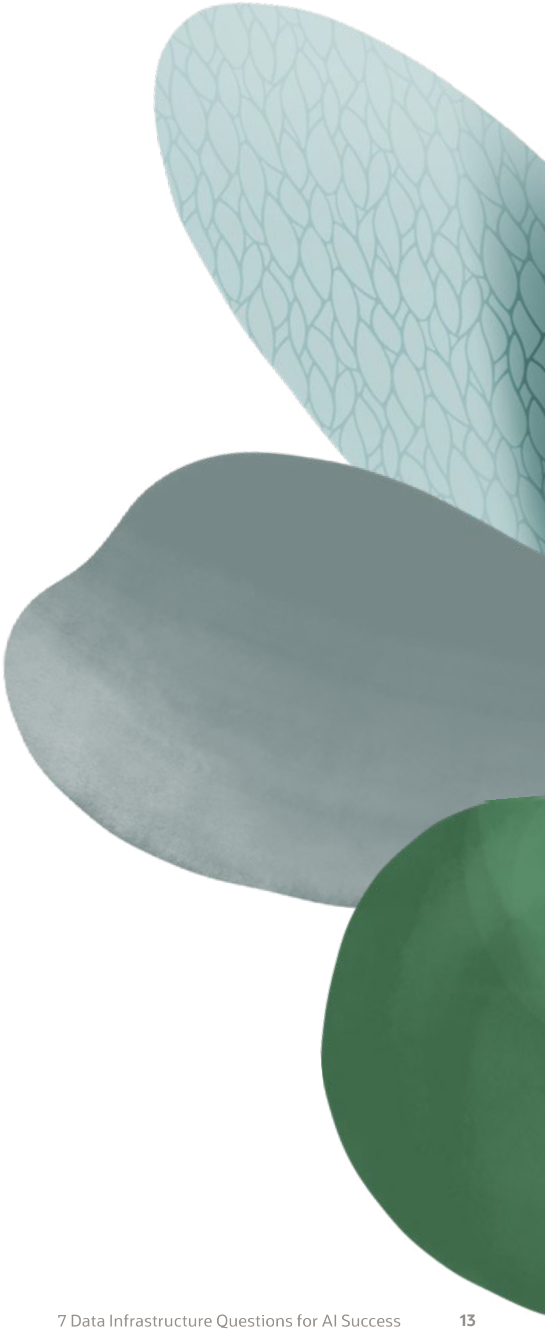
Next, understand your cloud providers' AI infrastructure services. Often you can get ready access to, or integration with, popular open source tools as well as GenAI foundation models and other necessary technologies.

Keep in mind that the open source community and hyperscalers are keen to help lower the costs of GenAI. The cloud is where you'll find the powerful GPUs and cluster networking needed to accelerate AI workloads as well as features such as auto-scaling, spot instances, reserved instances, optimized job scheduling, efficient workload distribution, multitenancy, and model optimization. These techniques maximize resource utilization—by matching compute power to the exact demands of the workload at any given time, companies pay only for what they use.

Fine-tuning is also getting, well, fine-tuned

A method known as T-Few can lower the training duration and the computational resources needed compared with conventional fine-tuning methods while still maintaining high accuracy.

Finally, you'll want to be able to prove your models and infrastructure are working at peak efficiency. You can monitor and optimize the inference process through your cloud provider's generative AI service or via efficient model serving frameworks, such as TensorFlow Serving or TorchServe.



7 Do we have executive backing for our AI plan? What groups will work with us?

Your AI strategy will most likely start with your current enterprise systems. Providers of CRM, HCM, ERP, and other core applications are embedding generative AI features and AI agents into common workflows. Think intelligent document recognition to help process supplier invoices faster and with much less manual work, generative narratives to bring a deeper understanding of reporting and analysis, and a wide range of intelligent automations in areas such as risk and performance analysis. These will light the way, but you'll want a broader strategy to bring AI to your unique workflows. This may involve the center of excellence discussed earlier.

Bringing generative AI services and AI agents into your business operations will require buy-in and resources from across the organization, including executives, department heads, IT, legal, compliance teams, and of course the people who will use the new technology. By working with stakeholders in each functional unit to develop the business case, focusing on productivity gains and competitive advantages, you can develop a coordinated approach.

Once you've settled on the AI workflows to start with, it's time to engage your broader IT and data science teams to understand what kind of architecture you'll need. Questions include: Where will the LLMs reside? Who will pay for them? How will the data flow—do we need new vector databases or RAG architectures? Do LLMs need to be fine-tuned for a particular department or task? Where will the data storage and compute power come from? Will our network design and other architectural decisions give us the low latency and high throughput we need for ongoing AI inferencing?

Prepare a detailed proposal with timelines, milestones, and resource requirements and an outline of risks, including data privacy, security and compliance, and cost considerations. Consider how your trusted cloud service providers can help.

> Keep in mind that no AI model will be effective without a steady flow of clean, well-ordered data

Therefore, consider your AI implementation an extension of your overall data strategy. To foster executive buy-in, teams will want to confirm their AI plans align closely with the organization's overall business strategy and data governance plan, without bringing in unnecessary complexity from moving data and adding management resources.

Oracle Solutions

When your data infrastructure needs to deliver fast, cost-effective fine-tuning and inferencing of the world's most advanced generative AI models, Oracle can help.

Oracle Cloud Infrastructure

Oracle AI Infrastructure

OCI provides comprehensive AI services and state-of-the-art generative AI innovations—all on a best-in-class AI infrastructure. You can choose prebuilt models or take advantage of a generative AI service that lets you choose from open source or proprietary LLMs and then fine-tune models and augment them with your own enterprise data. In addition, you have access to the industry-leading database management system.

Learn more

Oracle Database 23ai

Oracle Database 23ai is a multimodal database for transactions and analytics that provides a built-in vector database alongside JSON, relational, graph, and spatial data as well in-database machine learning. You can also do vector processing and run Oracle Database on uniquely optimized platforms that are available only in OCI and as OCI services in our multicloud partners' data centers.

Learn more

OCI Data Science and GenAI Services

OCI Data Science is a cloud service that lets data scientists collaboratively build, train, deploy, and manage machine learning (ML) models with your favorite open source frameworks, or they can leverage in-database ML. OCI's GenAI service is a fully managed service for integrating your choice of language models into a wide range of use cases.

Learn more

On OCI, you'll realize faster, more cost-effective fine-tuning, inferencing, and batch processing. OCI provides the scale and performance to run large AI workloads faster, without excessive compute costs.

How Oracle helps

Unique multicloud relationships can help ease data sharing.

Organizations worldwide use Oracle Database services running on OCI to quickly build and run applications on Oracle Exadata Database Service and take advantage of Oracle Autonomous Database capabilities. Multicloud relationships let organizations running applications in other hyperscale clouds get those same Oracle Database benefits with the simplicity, security, and low latency of a single operating environment.

Quickly build and run applications in AWS, using Exadata Database Service and Autonomous Database with services such as Amazon Bedrock; in Azure, using Exadata Database Service and Autonomous Database with services such as Azure OpenAI; and in Google Cloud, using Exadata Database Service and Autonomous Database with services such as Google Cloud's Vertex AI and Gemini foundation models.

Learn more

Connect with us

Call +1.800.ORACLE1 or visit oracle.com

Outside North America, find your local office at oracle.com/contact