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

Imagine the same LLMs that can explain quantum physics and summarize French novels gaining deep expertise in your company's unique operational data and knowledge base. 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 and take actions. IT resources are freed up and your organization is more data-driven.

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

The foundation for 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 a wide range of use cases. Success often depends on having enough power to handle complex AI operations—think GPUs interconnected by a cluster network that can achieve ultrahigh performance and microsecond latency. It also requires vector databases; a way to use AI agents for retrieval-augmented generation, or RAG, which combines your diverse enterprise knowledgebases with LLMs; and user-friendly AI agent interfaces.

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

What's Agentic AI?

Agentic AI refers to artificial intelligence that can understand and respond to information as well as actively pursue objectives.

Key characteristics of agentic AI

Proactive behavior
The AI can initiate actions rather than only 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 agentic AI is substantial. So are the required infrastructure investments.

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 ask as they work to deliver on the promise of AI now and into the future.

1 How will we benefit from a range of AI use cases, including some no one has 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 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 AI 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 ideas will likely yield creative AI use cases and a solid starting point for a data infrastructure plan.

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 involves taking a general-purpose model, such as Command from Cohere or Llama from Meta, and training it further on a smaller, domain-specific data set. This helps the model perform better on specific tasks because it has been adapted to the nuances and terminology of a domain, such as coding, finance, or healthcare. The 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 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 completed queries, alleviating a potential source of data leaks.

2 How will we make all of our unstructured and semistructured data available for AI?

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

Data for fine-tuning: Fine-tuning an AI model for a specific task can require showing it new discipline-specific data. For instance, if you're tuning a generative AI model to produce medical reports, you need a data set of high-quality, relevant medical reports to help the model learn the appropriate 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, relational data, and semistructured data. It needs to efficiently apply vector embeddings to data, store them 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 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 record every customer call, using AI to analyze sentiment and generate summaries. No more indecipherable scribbling as salespeople try to engage customers 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 more accurate and efficient.

One key to success: A converged, multimodal database that supports 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 vectors, without IT needing to integrate disparate silos. This can lead to richer results, as AI models evaluate 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 data provided to your generative AI tools is a core infrastructure requirement—across model fine-tuning, RAG workflows, and day-to-day operations.

Depending on your needs, data for your AI system can be anonymized, encrypted, or masked. Access to data used in AI outputs must be tightly controlled by considering the requester's role and employing a multifactor authentication process. 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 more than 280 organizations collaborating with the National Institute of Standards and Technology (NIST) in its NIST AI Consortium. One 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 aim to develop guidelines, tools, methods, protocols, and best practices to help the community develop and deploy AI securely.

AI architecture: Multimodal database versus multiple single-use databases

With a multimodal database, you typically don't need extensive ETL and data movement when adopting generative AI. A single data platform can support many diverse AI projects and workloads, significantly reducing complexity and management overhead while limiting risks from moving data. A multimodal database doesn't require data to be in one format or model; it allows different data architectures for each type of application.

4 How will teams collaborate on our AI 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. Giving your workers AI tools increases the likelihood that AI workflows become part of your company's strategy—and not isolated projects.

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

Learn how to build an AI CoE in your company

A core technology for collaboration is a data science platform that eencourages people to work together on the technical steps involved in fine-tuning and deploying models. Your hyperscale cloud provider will have an AI data platform for working with foundation models from Cohere, Meta, and other suppliers. These platforms give data scientists, IT teams, and developers 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. This lets business units build 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 options for fully managed services. Oracle AI Data Platform brings together AI models and governed enterprise data so developers can rapidly and easily build AI agents and applications that business users can then apply to operational workflows.

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

Your organization may be evaluating Google Gemini or Microsoft's Copilot, or hosting open source models in AWS or Cohere's foundation models in Oracle Cloud Infrastructure (OCI). Agreements between Oracle and other hyperscalers expand options for companies whose data governance strategies are built around Oracle AI Database technology, whether deployed on-premises or in the cloud. You can embrace 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 AI Database services in OCI and 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 use popular data science platforms from major cloud providers to make data sets clean and relevant to a task.

It's important to understand available AI infrastructure services. Cloud providers offer access to, or integration with, popular open source tools, AI foundation models, and supporting technologies.

Keep in mind that the open source community and hyperscalers can help lower the costs of AI. Cloud environments provide the powerful GPUs and high performance networking needed for AI workloads, along with capabilities 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 time and 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 operate efficiently. You can monitor and optimize the inference process through your cloud provider's generative AI service or via model-serving frameworks, such as TensorFlow Serving or TorchServe.

7 Do we have executive backing for our AI plan? Which 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 capabilities and AI agents into common workflows. Examples include intelligent document recognition to help process supplier invoices faster and with much less manual work, generative narratives for a deeper understanding of reporting and analysis, and a wide range of intelligent automations in areas such as risk and performance analysis. These capabilities provide a starting point, but you'll want a broader strategy to apply AI across your unique workflows. This may involve establishing a center of excellence, as discussed earlier.

Implementing generative AI services and AI agents into your business operations will require buy-in and resources from multiple groups, 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 a business case, focusing on productivity gains and competitive advantages, you can develop a coordinated approach.

Once initial AI workflows are defined, IT and data science teams must determine architecture requirements. Questions to address 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, resource requirements, and an outline of risks, including data privacy, security, compliance, and costs. Consider how your cloud service providers can help.

> Keep in mind that no AI model will be effective without access to a steady flow of clean, well-structured 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 must support fast, cost-effective fine-tuning and inference for advanced generative AI models, Oracle can help.

Oracle Cloud Infrastructure

Oracle AI Data Platform

Oracle AI Data Platform brings together all types of data—structured, unstructured, batch, and real time—across your enterprise into an open and connected platform, laying the foundation for trusted and AI-ready data pipelines. It provides an environment where developers can easily build AI applications and agents and deploy them on Oracle's AI infrastructure, so you can scale innovation and intelligence everywhere your business operates.

Learn more

Oracle AI infrastructure

Use Oracle's AI infrastructure to run the most demanding AI workloads, including frontier model training, inference, agentic AI, scientific computing, and recommender systems, anywhere in our distributed cloud. You can even take advantage of OCI Supercluster for GPU performance at zettascale.

Learn more

Oracle AI Database

Oracle AI Database is the only database that natively architects AI into your data everywhere. Reduce complexity, risk, and cost by converging every workload into a single unified environment, backed by stock exchange–grade security and scale. Unique multicloud relationships dramatically simplifies data sharing with AWS, Azure, and Google Cloud.

Learn more

OCI Enterprise AI

Build and deploy production-ready AI agents across data sources using open standards, open source frameworks, and managed access to leading models with zero-data-retention endpoints, backed by sovereign AI hosting and enterprise-grade governance for secure, scalable deployment.

Learn more

How Oracle helps

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

Learn more

Connect with us

Call +1.800.ORACLE1 or visit oracle.com

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