10 tips for building a data foundation for generative artificial intelligence

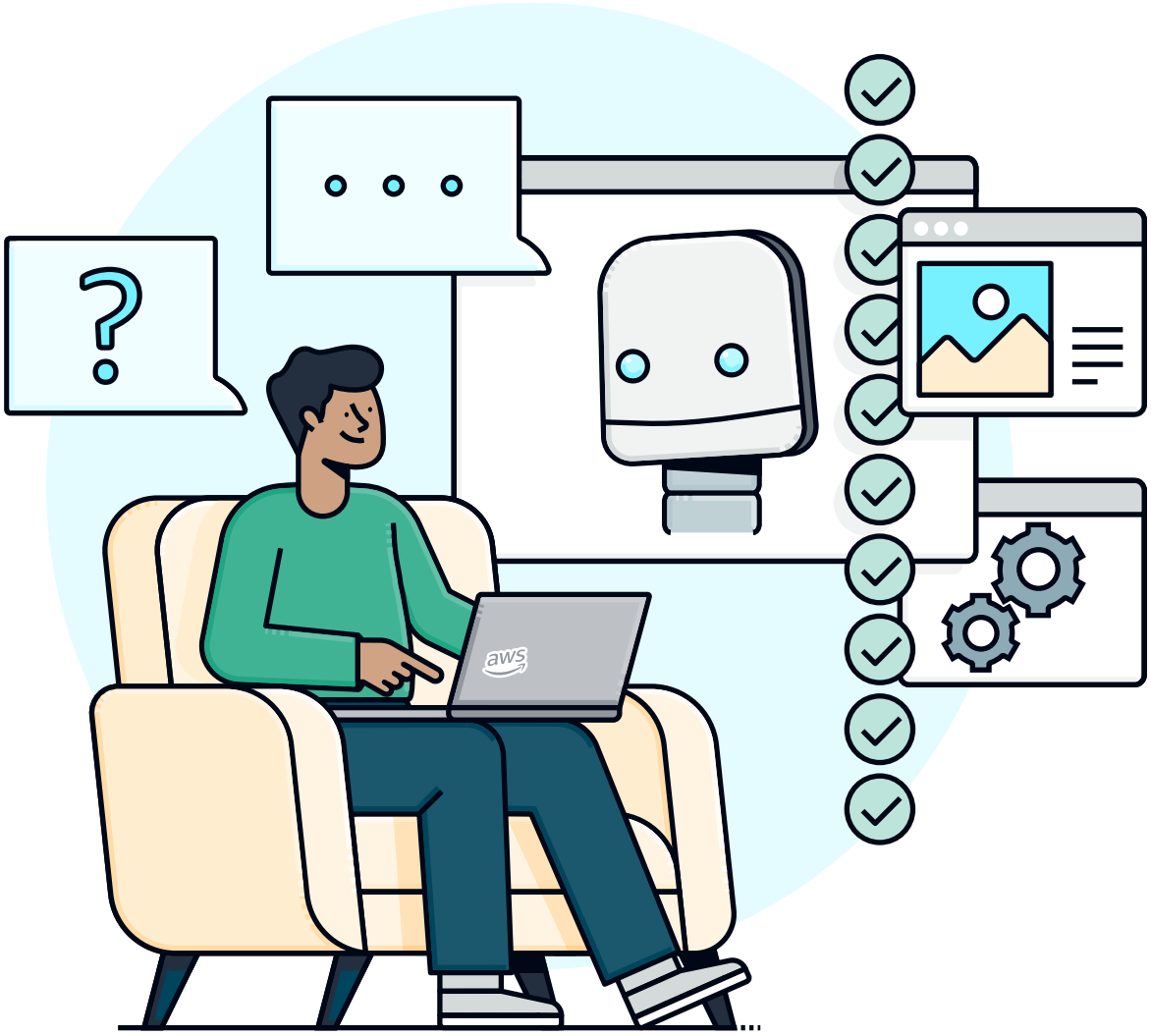


Table of contents

Introduction

Generative artificial intelligence (AI) is transforming how organizations approach problems, create new offerings, and interact with customers. When you want to increase your productivity or build differentiated generative AI applications that deliver unique value for your business, having access to a variety of high-quality, well-governed data is crucial. Providing high-quality, relevant data to generative AI models requires more than only gathering and storing data. It requires implementing an end-to-end data foundation to efficiently understand, curate, and protect your information for analytics and generative models. This strategy needs to scale with your business and adapt to evolving technologies.

Popular generative AI use cases:

Text generation: Create new pieces of original content, such as blog posts, social media posts, and webpage copy

Virtual assistants: Build assistants that understand user requests, automatically break down tasks, engage in dialogue to collect information, and take action to fulfill requests

Text and image search: Search and synthesize relevant information to answer questions and provide recommendations from a large corpus of text and image data

Text summarization: Get concise summaries of long documents—such as articles, reports, research papers, technical documentation, and even books—to quickly and effectively extract important information

Image generation: Quickly create realistic and visually appealing images for ad campaigns, websites, presentations, and more

Why generative AI is a forcing function for data practices

Generative AI, by design, is data hungry. Large language models (LLMs) require expansive datasets to inform the generation of new text, images, audio, and more. To make the most of generative AI technologies such as LLMs and natural language processing, your data architecture must support new workloads with large volumes of complex, transactional, and contextual data. It also needs to address compliance, responsibility, and bias concerns with technologies that learn from data. With generative AI, data practices can no longer be an afterthought—they must be central to how your organization innovates and operates.

Executive sponsorship

To get ongoing support for data initiatives and avoid the risk of losing funding, executives must tie the efforts to business outcomes. Highlight how data directly enables major projects and helps achieve the goals of the funded business initiatives. With executive sponsorship, you can more easily align data practices across departments—and across business initiatives—and justify needed investments. Keep in mind, you need executive sponsorship for the people and process components too.

Work backwards from funded business initiatives

Before implementing new tools or technologies, evaluate your current and future data needs by analyzing your strategic plans and funded business initiatives. Start with the vision of each business outcome and work backwards to define the data foundation required for the initiative. Then identify which components can be accelerated by generative AI and which will require additional data capabilities. Instead of proposing the value of a data foundation on its own, work backwards from a business initiative that includes a generative AI component.

Risk of getting left behind competitively

Without a data-fueled transformation, your business risks falling behind rapidly changing customer and industry demands. By modernizing your data foundation, you can explore new markets and revenue streams before competitors. You can also retain your top employees interested in cutting-edge work, such as vector engines and LLMs.

Build a data foundation for generative AI

To enable generative AI, you need a foundation that includes a comprehensive, integrated set of data services for all workloads, use cases, and types of data—and tools to govern that data. Amazon Web Services (AWS) offers a broad set of integrated capabilities, from databases and data lakes to analytics and machine learning. The services are purpose-built for scaling needs, with high performance, security, and dependability guaranteed by service level agreements.

Tip 1:

Find the best data for generative AI

Although any data provides value, high-quality sources that align with business initiatives yield the most relevant results. Consider traditional data-quality factors such as accuracy, completeness, and relevancy when assessing potential AI training data. Automate choices based on the metadata you've captured. Prioritize common data issues that degrade model outcomes.

Tip 2:

Pay attention to metadata

Metadata provides rich context on your data that generative models can use to improve results. Ensure your practices and tools support standardized, insightful metadata capture from ingest to long-term storage. Manual metadata capture is very time-consuming, so put a priority on using new generative AI capabilities on your own data. Benefits include more uniform analytics output, superior data discoverability, and better governance controls.

Tip 3:

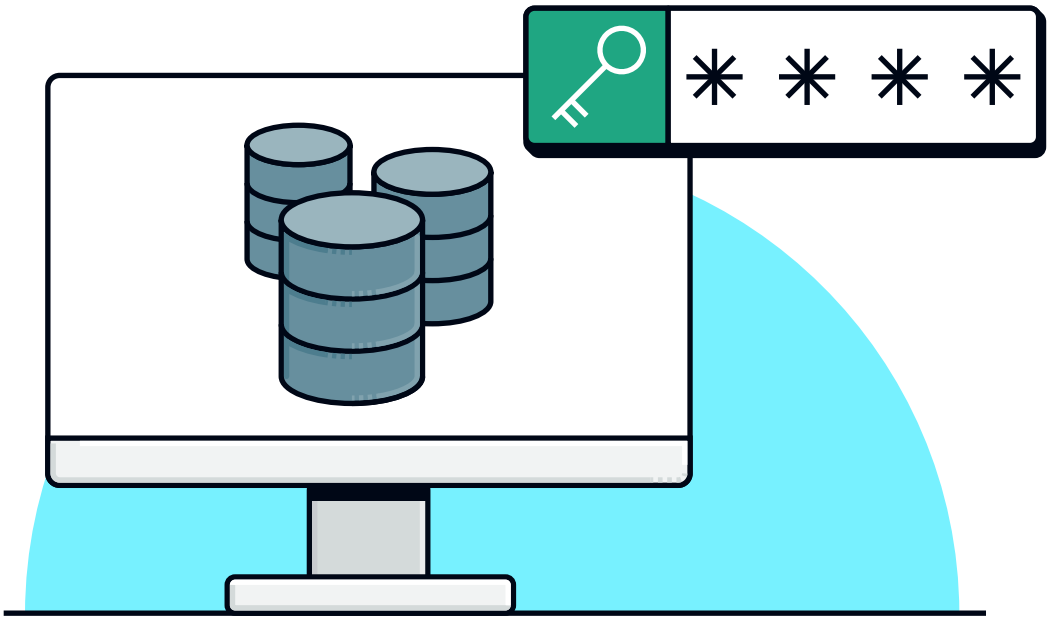
Understand context and bias in data

All data reflects real-world biases that generative models can absorb and amplify. Review your data sources and metadata practices for biases related to protected attributes. Mitigation strategies help governance teams address biases, and the technical teams build responsibility into models from the start. Leaders set the tone for accountability throughout development and production use.

Tip 4:

Automate data access controls for models

Granting generative models access to the right data sources requires tightly governed access. Your models should subscribe to the right datasets—much like a traditional data consumer—so the governance system has complete visibility. Similarly, fine-grained access policies ensure transformations, embeddings, and other derivatives respect privacy and compliance rules.



Tip 5:

Minimize data movement and reprocessing

Operations such as extract, transform, and load (ETL) consume time and resources that are better spent on innovation. AWS investments in zero-ETL and the built-in vector capabilities in our popular databases reduce data friction and movement, so more of your effort can support innovation. By optimizing data workflows for minimal movement, you can reduce preprocessing and duplication between repositories and services.

Tip 6:

Govern features centrally

Versioned, discoverable features fuel repeatable training and fair, accountable decision making. Govern edits, annotation practices, and model training loops online or through lifecycle actions. Traceability boosts transparency and prevents workforce bottlenecks.

Tip 7:

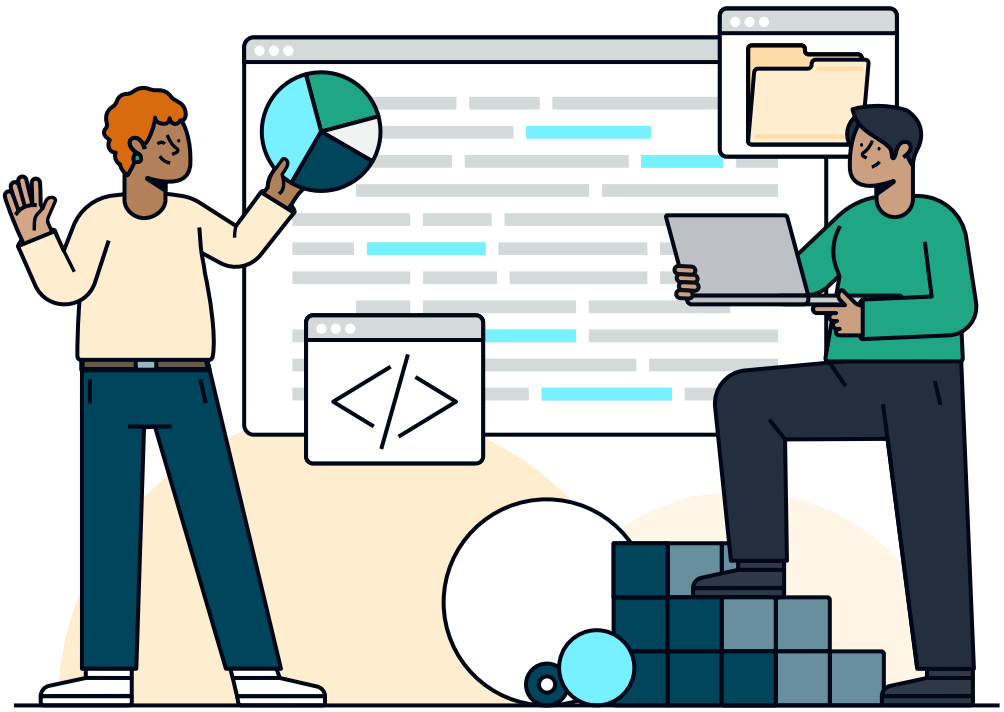
Align machine learning operations and governance

Building, testing, and deploying models into production requires processes and technology to enable efficiency, reliability, and speed. Many of these processes are also used in a data governance practice. Use the synergies of machine learning operations (MLOps) and data governance to optimize your processes and reporting.

Tip 8:

Document model decisions

Technical choices during generative AI development have an impact on real users, yet reasoning behind them remains opaque without documentation. There's a wide variety of commercial and open-source foundation models available for various uses, and they're each good at different things. Applying governance to the selection process—that is, guiding which model to use under what circumstances—is very helpful, especially when we need to chain multiple models together for solutions. Transparency builds the accountability that's essential for responsible, business-wide adoption.



Tip 9:

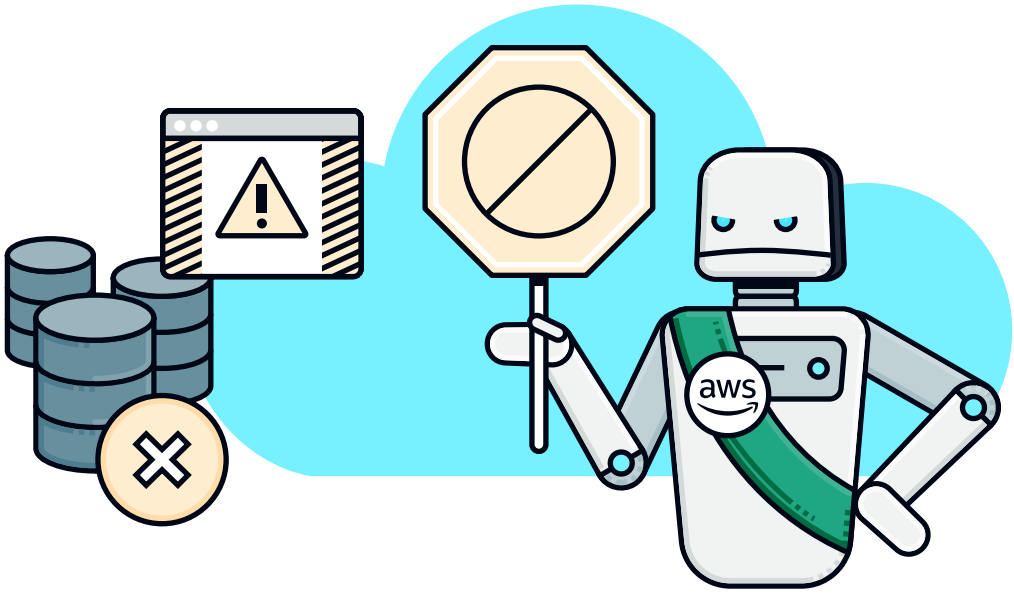
Quality still matters, but standards evolve

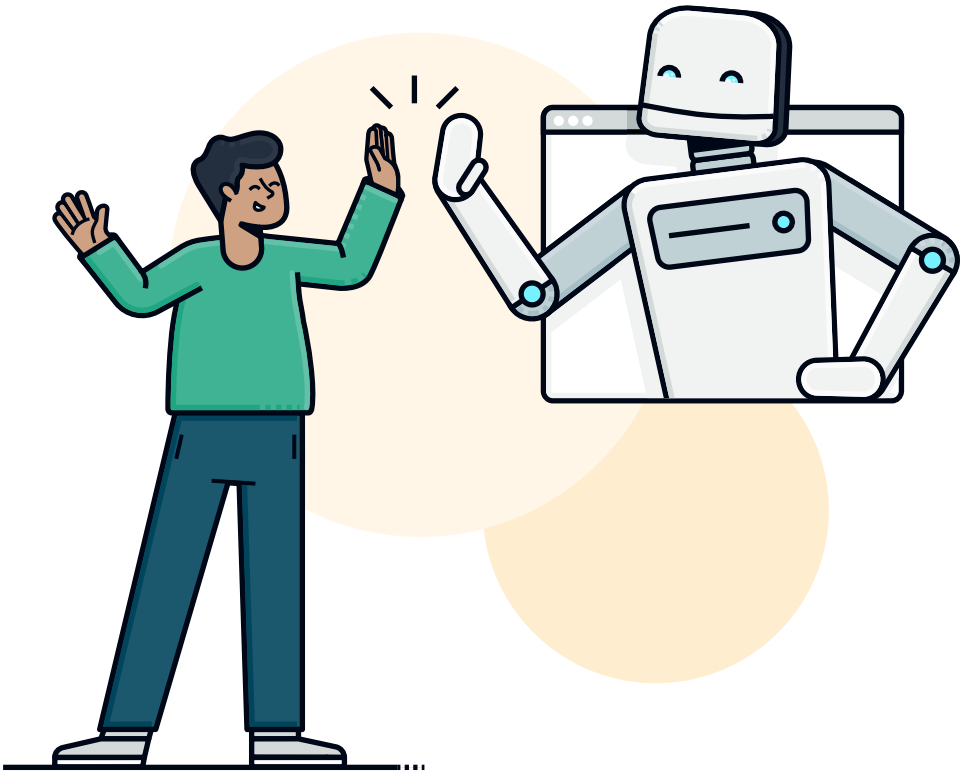
In analytics and AI, quality definitions broaden from accuracy and validity alone. But when the output is generated content, make sure you define what good enough looks like. You may find it easier to define the converse—what absolutely bad generative results look like. Keep track of these bad results and use them as test cases as models move into production. Partner model training with behavioral analysis for optimized recommendations and nudges.

Tip 10:

Control access to data used for Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) is one approach you can use to redirect an LLM to retrieve relevant information from authoritative, predetermined knowledge sources. RAG helps enhance generative AI models with factual knowledge-retrieval capabilities, resulting in more accurate, evidence-based outputs. Using a RAG approach may give you greater control over the generated text output. However, you need to control which data is provided in a RAG approach to make sure sensitive or lower-quality data is not used.





Conclusion

Data products and generative AI applications today deliver new business benefits certain to multiply over time. Yet for any organization, innovation hinges on capabilities for accessing, understanding, and responsibly governing all available information. You can build a flexible data foundation with AWS. An end-to-end data foundation built on AWS equips your teams to meet customer and industry changes today while developing solutions for the future. In this way, you'll be better prepared for the next generative AI–style moment.

Learn how you can build a data foundation for generative AI ›