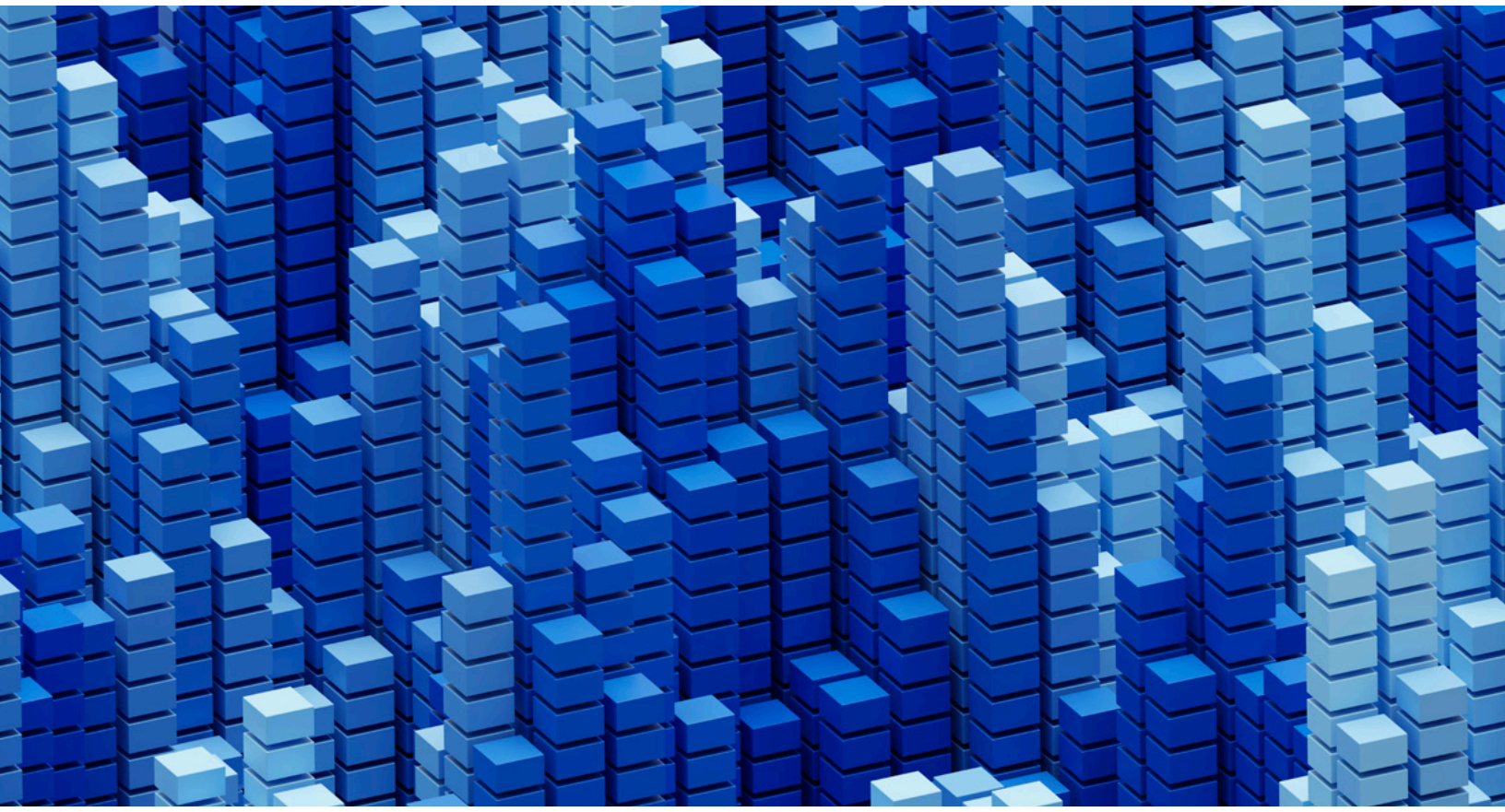
McKinsey Digital

The data dividend: Fueling generative AI

Data leaders should consider seven actions to enable companies to scale their generative AI ambitions.

This article is a collaborative effort by Joe Caserta, Holger Harreis, Kayvaun Rowshankish, Nikhil Srinidhi, and Asin Tavakoli, representing views from McKinsey Digital.



September 2023

If your data isn't ready for generative AI, your business isn't ready for generative AI.

Our latest research estimates that generative AI could add the equivalent of $2.6 trillion to $4.4 trillion in annual economic benefits across 63 use cases.[1] Pull the thread on each of these cases, and it will lead back to data. Your data and its underlying foundations are the determining factors to what's possible with generative AI.

That's a sobering proposition for most chief data officers (CDOs), especially when 72 percent of leading organizations note that managing data is already one of the top challenges preventing them from scaling AI use cases.[2] The challenge for today's CDOs and data leaders is to focus on the changes that can enable generative AI to generate the greatest value for the business.

The landscape is still rapidly shifting, and there are few certain answers. But in our work with more than a dozen clients on large generative AI data programs, discussions with about 25 data leaders at major companies, and our own experiments in reconfiguring data to power generative AI solutions, we have identified seven actions that data leaders should consider as they move from experimentation to scale:

1. **Let value be your guide.** CDOs need to be clear about where the value is and what data is needed to deliver it.

2. **Build specific capabilities into the data architecture to support the broadest set of use cases.** Build relevant capabilities (such as vector databases and data pre- and post-processing pipelines) into the existing data architecture, particularly in support of unstructured data.

3. **Focus on key points of the data life cycle to ensure high quality.** Develop multiple interventions—both human and automated—into the data life cycle from source to consumption to ensure the quality of all material data, including unstructured data.

4. **Protect your sensitive data, and be ready to move quickly as regulations emerge.** Focus on securing the enterprise's proprietary data and protecting personal information while actively monitoring a fluid regulatory environment.

5. **Build up data engineering talent.** Focus on finding the handful of people who are critical to implementing your data program, with a shift toward more data engineers and fewer data scientists.

6. **Use generative AI to help you manage your own data.** Generative AI can accelerate existing tasks and improve how they're done along the entire data value chain, from data engineering to data governance and data analysis.

7. **Track rigorously and intervene quickly.** Invest in performance and financial measurement, and closely monitor implementations to continuously improve data performance.

1. Let value be your guide

In determining a data strategy for generative AI, CDOs might consider adapting a quote from President John F. Kennedy: "Ask not what your business can do for generative AI; ask what generative AI can do for your business." Focus on value is a long-standing principle, but CDOs must particularly rely on it to counterbalance the pressure to "do something" with generative AI. To provide this focus on value, CDOs will need to develop a clear view of the data implications of the business's overall approach to generative AI, which will play out across three archetypes:

— **Taker:** a business that consumes preexisting services through basic interfaces such as APIs. In this case, the CDO will need to focus on making quality data available for generative AI models and subsequently validating the outputs.

[1] "The economic potential of generative AI: The next productivity frontier," McKinsey, June 14, 2023.
[2] McKinsey Data & AI Summit 2022.

— **Shaper:** a business that accesses models and fine-tunes them on its own data. The CDO will need to assess how the business's data management needs to evolve and what changes to the data architecture are needed to enable the desired outputs.

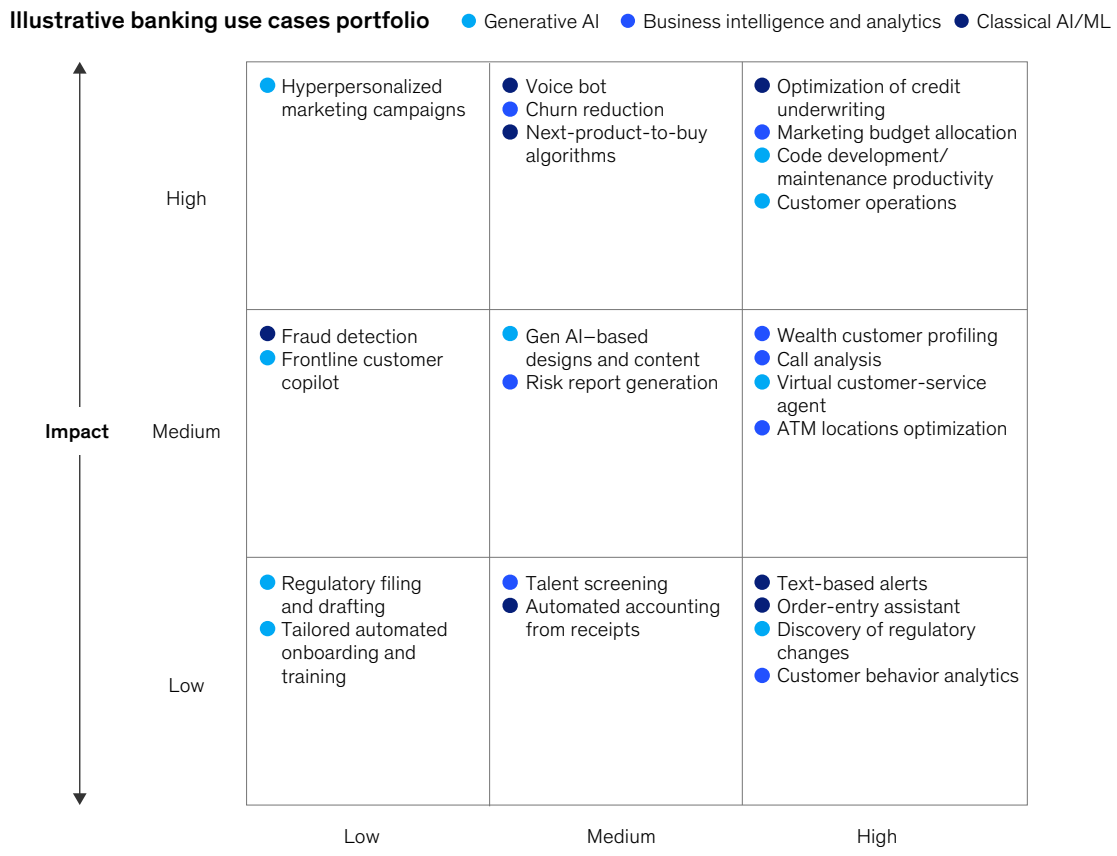
— **Maker:** a business that builds its own foundational models. The CDO will need to develop a sophisticated data labeling and tagging strategy, as well as make more significant investments.

The CDO has the biggest role to play in supporting the Shaper approach, since the Maker approach is currently limited to only those large companies willing to make major investments and the Taker approach essentially accesses commoditized capabilities. One key function in driving the Shaper approach is communicating the trade-offs needed to deliver on specific use cases and highlighting those that are most feasible. While hyperpersonalization, for example, is a promising generative AI use case, it requires clean customer data, strong guardrails for data protection, and pipelines to access multiple data sources. The CDO should also prioritize initiatives that can provide the broadest benefits to the business, rather than simply support individual use cases.

As CDOs help shape the business's approach to generative AI, it will be important to take a broad view on value. As promising as generative AI is, it's just one part of the broader data portfolio (Exhibit 1). Much of the potential value to a business comes from traditional AI, business intelligence, and machine learning (ML). If CDOs find themselves spending

Exhibit 1

Take a portfolio view on value.



90 percent of their time on initiatives related to generative AI, that's a red flag.

2. Build specific capabilities into the data architecture to support the broadest set of use cases

The big change when it comes to data is that the scope of value has gotten much bigger because of generative AI's ability to work with unstructured data, such as chats, videos, and code. This represents a significant shift because data organizations have traditionally had capabilities to work with only structured data, such as data in tables. Capturing this value doesn't require a rebuild of the data architecture, but the CDO who wants to move beyond the basic Taker archetype will need to focus on two clear priorities.

The first is to fix the data architecture's foundations. While this might sound like old news, the cracks in the system a business could get away with before will become big problems with generative AI. Many of the advantages of generative AI will simply not be possible without a strong data foundation. To determine the elements of the data architecture on which to focus, the CDO is best served by identifying the fixes that provide the greatest benefit to the widest range of use cases, such as data-handling protocols for personally identifiable information (PII), since any customer-specific generative AI use case will need that capability.

The second priority is to determine which upgrades to the data architecture are needed to fulfill the requirements of high-value use cases. The key issue here is how to cost effectively manage and scale the data and information integrations that power generative AI use cases. If they are not properly managed, there is a significant risk of overstressing the system with massive data compute activities, or of teams doing one-off integrations, which increase complexity and technical debt. These issues are further complicated by the business's cloud profile, which means CDOs must work closely with IT leadership to determine compute, networking, and service use costs.

In general, the CDO will need to prioritize the implementation of five key components of the data architecture as part of the enterprise tech stack (Exhibit 2):

- **Unstructured data stores:** Large language models (LLMs) primarily work with unstructured data for most use cases. Data leaders will need to map out all unstructured data sources and establish metadata tagging standards so models can process the data and teams can find the data they need. CDOs will need to further upgrade the quality of data pipelines and establish standards for transparency so that it's easy to track the source of an issue to the right data source.

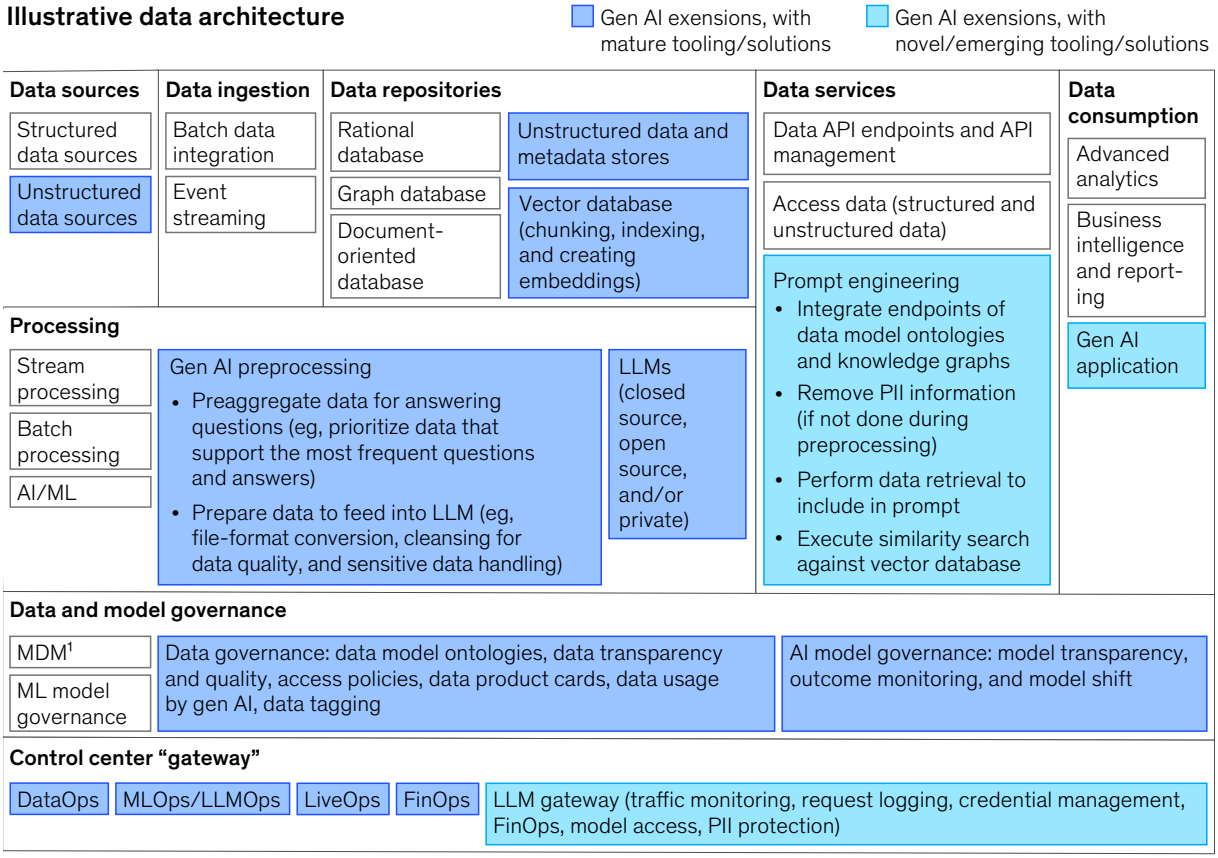
- **Data preprocessing:** Most data will need to be prepped—for example, by converting file formats and cleansing for data quality and the handling of sensitive data—so that generative AI can use the data. Preprocessed data is most often used to build prompts for generative AI models. To speed up performance, CDOs need to standardize the handling of structured and unstructured data at scale, such as ways to access underlying systems, and prioritize (or "preaggregate") the data that supports the most frequent questions and answers.

- **Vector databases:** Vectorization is a way to prioritize content and create "embeddings" (numerical representations of text meanings) in order to streamline access to context, the complementary information generative AI needs to provide accurate answers. Vector databases allow generative AI models to access just the most relevant information. Instead of providing a thousand-page PDF, for example, a vector database provides only the most relevant pages. In many cases, companies don't need to build vector databases to begin working with generative AI. They can often use existing NoSQL databases to start.

- **LLM integrations:** More-sophisticated generative AI uses require interactions with

Exhibit 2

Upgrades are needed within the existing data architecure to enable generative AI.



[1]Master data management.

McKinsey & Company

multiple systems, which creates significant challenges in connecting LLMs. Several frameworks, many of which are open source, can help facilitate these integrations (for example, LangChain or various hyperscaler offerings, such as Semantic Kernel for Azure, Bedrock for AWS, or Vertex AI for Google Cloud). CDOs will need to set guidelines for choosing which frameworks to use, define prompt templates that can be readily customized for specific purposes, and establish standardized integration patterns for how LLMs interface with source data systems.

— **Prompt engineering:** Effective prompt engineering (the process of structuring questions in a way that elicits the best response from generative AI models) relies on context. Context can be determined only from existing data and information across structured and unstructured sources. To improve output, CDOs will need to manage integration of knowledge graphs or data models and ontologies (a set of concepts in a domain that shows their properties and the relations between them) into the prompt. Since CDOs will not have ownership of many data repositories across the business, they

need to set standards and prequalify sources to ensure the data that is fed into the models follows specific protocols (for example, exposing a knowledge graph API to easily provide entities and relationships).

3. Focus on key points of the data life cycle to ensure high quality

Data quality has always been an important issue for CDOs. But the scale and scope of data that generative AI models rely on has made the "garbage in/garbage out" truism much more consequential and expensive, as training a single LLM can cost millions of dollars.[3] One reason pinpointing data quality issues is much more difficult in generative AI models than in classical ML models is because there's so much more data and much of it is unstructured, making it difficult to use existing tracking tools.

CDOs need to do two things to ensure data quality: extend their data observability programs[4] for generative AI applications to better spot quality issues, such as by setting minimum thresholds for unstructured content to be included in generative AI applications; and develop interventions across the data life cycle to fix the issues teams find, mainly in four areas:

— **Source data:** Expand the data quality framework to include measures relevant for generative AI purposes (such as bias). Ensure high-quality metadata and labels for structured and unstructured data, and regulate access to sensitive data (for example, base access on roles).

— **Preprocessing:** Ensure data is consistent and standardized and adheres to ontologies and established data models. Detect outliers and apply normalizations. Automate PII data management, and put in place guidelines for whether data should be ignored, held, redacted, quarantined, removed, masked, or synthesized.

— **Prompt:** Evaluate, measure, and track the quality of the prompt. Include high-quality metadata and lineage transparency for structured and unstructured data in the prompt.

— **Output from LLM:** Establish the necessary governance procedures to identify and resolve incorrect outputs, and use "human in the loop" to review and triage output issues. Ultimately, elevate the role of individual employees by training them to critically evaluate model outputs and be aware of the quality of input data. Supplement with an automated monitoring-and-alert capability to identify rogue behaviors.

4. Protect your sensitive data, and be ready to move quickly as regulations emerge

Some 71 percent of senior IT leaders believe generative AI technology is introducing new security risk to their data.[5] Much has been written about security and risk when it comes to generative AI, but CDOs needs to consider the data implications in three specific areas:

— **Identify and prioritize security risks to the enterprise's proprietary data.** CDOs need to assess the broad risks associated with exposing the business's data, such as the potential exposure of trade secrets when confidential and proprietary code is shared with generative AI models, and prioritize the greatest threats. Much existing data protection and cybersecurity governance can be extended to address specific generative AI risks—for example, by adding pop-up reminders whenever an engineer wants to share data with a model or by running automated scripts to ensure compliance.

— **Manage access to PII data.** CDOs need to regulate how data is detected and treated in the context of generative AI. They need to set up systems that incorporate protection tools and human interventions to ensure PII

[3] Urian B., "NVIDIA announces $9.6 million drop in cost when using its GPUs for AI LLM training," Tech Times, May 29, 2023.
[4] Data observability programs consist of mechanisms for understanding the health and performance of the data within systems.
[5] "Top generative AI statistics for 2023," Salesforce, September 2023.

data is removed during data preprocessing and before it's used on an LLM. Using synthetic data (through data fabricators) and nonsensitive identifiers can help.

— **Track the expected surge of regulations closely.** Generative AI has acted as a catalyst to rapid movement among governments to enact new regulations, such as the European Union's AI Act, which is setting a wide array of new standards, such as having companies publish summaries of copyrighted data used for training an LLM. Data leaders must stay close to the business's risk leaders to understand new regulations and their implications for data strategy, such as the need to "untrain" models that use regulated data.

5. Build up data engineering talent

As enterprises increasingly adopt generative AI, CDOs will have to focus on the implications for talent. Some coding tasks will be done by generative AI tools—41 percent of code published on GitHub is written by AI.[6] This requires specific training on working with a generative AI "copilot"—a recent McKinsey study showed that senior engineers work more productively with a generative AI copilot than do junior engineers.[7] Data and AI academies need to incorporate generative AI training tailored to specific expertise levels.

CDOs will also need to be clear about what skills best enable generative AI. Companies need people who can integrate data sets (such as writing APIs connecting models to data sources), sequence and chain prompts, wrangle large quantities of data, apply LLMs, and work with model parameters. This means that CDOs should focus more on finding data engineers, architects, and back-end engineers, and less on hiring data scientists, whose skills will be increasingly less critical as generative AI allows people with less advanced technical capabilities to use natural language in doing basic analysis.

In the near term, talent will remain in shorter supply, and we project that the talent gap will increase further in the near future,[8] creating more incentives for CDOs to build up their training programs.

6. Use generative AI to help you manage data

Data leaders have a huge opportunity to harness generative AI to improve their own function. In our analysis, eight primary use cases have emerged along the entire data value chain where generative AI can both accelerate existing tasks and improve how tasks are performed (Exhibit 3).

Many vendors are already rolling out products, requiring CDOs to identify the capabilities for which they can rely on vendors and which they should build themselves. One rule of thumb is that for data governance processes that are unique to the business, it's better to build your own tool. Note that many tools and capabilities are new and may work well in experimental environments but not at scale.

7. Track rigorously and intervene quickly

There are more unknowns than knowns in the generative AI world today, and companies are still learning their way forward. It is therefore crucial for CDOs to set up systems to actively track and manage progress on their generative AI initiatives and to understand how well data is performing in supporting the business's goals.

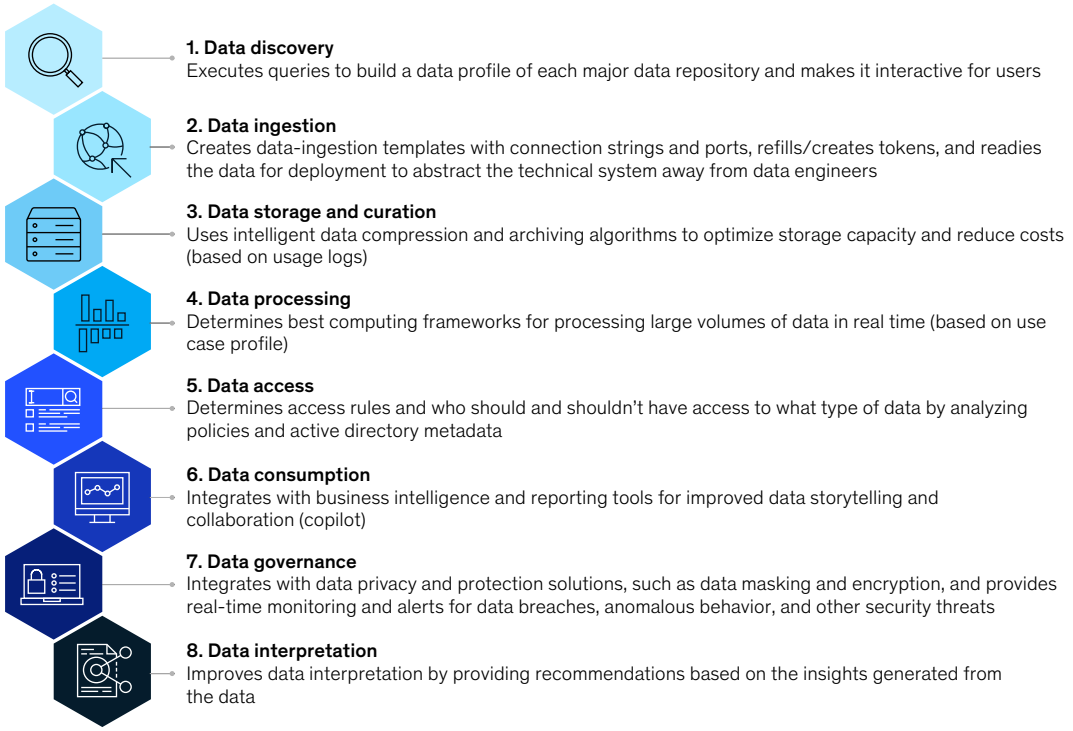
In practice, effective metrics are made up of a set of core KPIs and operational KPIs (the underlying activities that drive KPIs), which help leaders track progress and identify root causes of issues.

[6] Jose Antonio Lanz, "Stability AI CEO: There will be no (human) programmers in five years," Decrypt, June 3, 2023.
[7] "Unleashing developer productivity with generative AI," McKinsey, June 27, 2023.
[8] Michael Chui, Mena Issler, Roger Roberts, and Lareina Yee, "McKinsey Technology Trends Outlook 2023," McKinsey, July 20, 2023.

Exhibit 3

Generative AI opportunities exist to improve the entire date value chain.

Generative AI use cases along data value chain



A core set of KPIs should include the following:

— cost of additional components, such as vector databases and consumption of LLMs as a service

— additional revenue that is enabled by the integration of specific data sources with generative AI application workflows

— time-to-market to develop a generative AI–powered application that requires access to internal data

— end-user satisfaction with how the data has improved the performance and quality of the application

Operational KPIs should include tracking which data are being used most, how models are performing, where data quality is poor, how many requests are being made against a given data set, and which use cases are generating the most activity and value.

This information is critical in providing a fact base for leadership to not just track progress but also make rapid adjustments and trade-off decisions against other initiatives in the CDO's broader portfolio. By knowing which data sources are most used for high-value models, for example, the CDO can prioritize investments to improve data quality at those sources.

Effective investment, budgeting, and reallocation will depend on CDOs developing a FinOps-like capability to manage the entire new cost structure growing around generative AI. CDOs will need to track a new range of costs, including the number of generative AI model requests, API consumption

charges from vendors (both quantity and size of calls), and compute and storage charges from cloud providers. With this information, the CDO can determine how best to optimize costs, such as routing requests by priority level or moving certain data to the cloud to cut down on networking costs.

The value of these metrics is only as great as the degree to which CDOs act on them. CDOs will need to establish data-performance metrics that can be reviewed in near real time and protocols to make rapid decisions. Effective data governance programs should remain in place but be extended to incorporate generative AI–related decisions.

Data cannot be an afterthought in generative AI. Rather, it is the core fuel that powers the ability of a business to capture value from generative AI. But businesses that want that value cannot afford CDOs who merely manage data; they need CDOs who understand how to use data to lead the business.

Joe Caserta is a partner in McKinsey's New York office, where **Kayvaun Rowshankish** is a senior partner; **Holger Harreis** is a senior partner in the Düsseldorf office, where **Asin Tavakoli** is a partner; and **Nikhil Srinidhi** is an associate partner in the Berlin office.

The authors wish to thank Sven Blumberg, Stephanie Brauckmann, Carlo Giovine, Jonas Heite, Vishnu Kamalnath, Simon Malberg, Rong Parnas, Bruce Philp, Adi Pradhan, Alex Singla, Saravanakumar Subramaniam, Alexander Sukharevsky, and Kevin-Morris Wigand for their contributions to this article.