The Chief Data Officer's guide to AI transformation

Guarantee your organization's readiness as AI continues to mature.

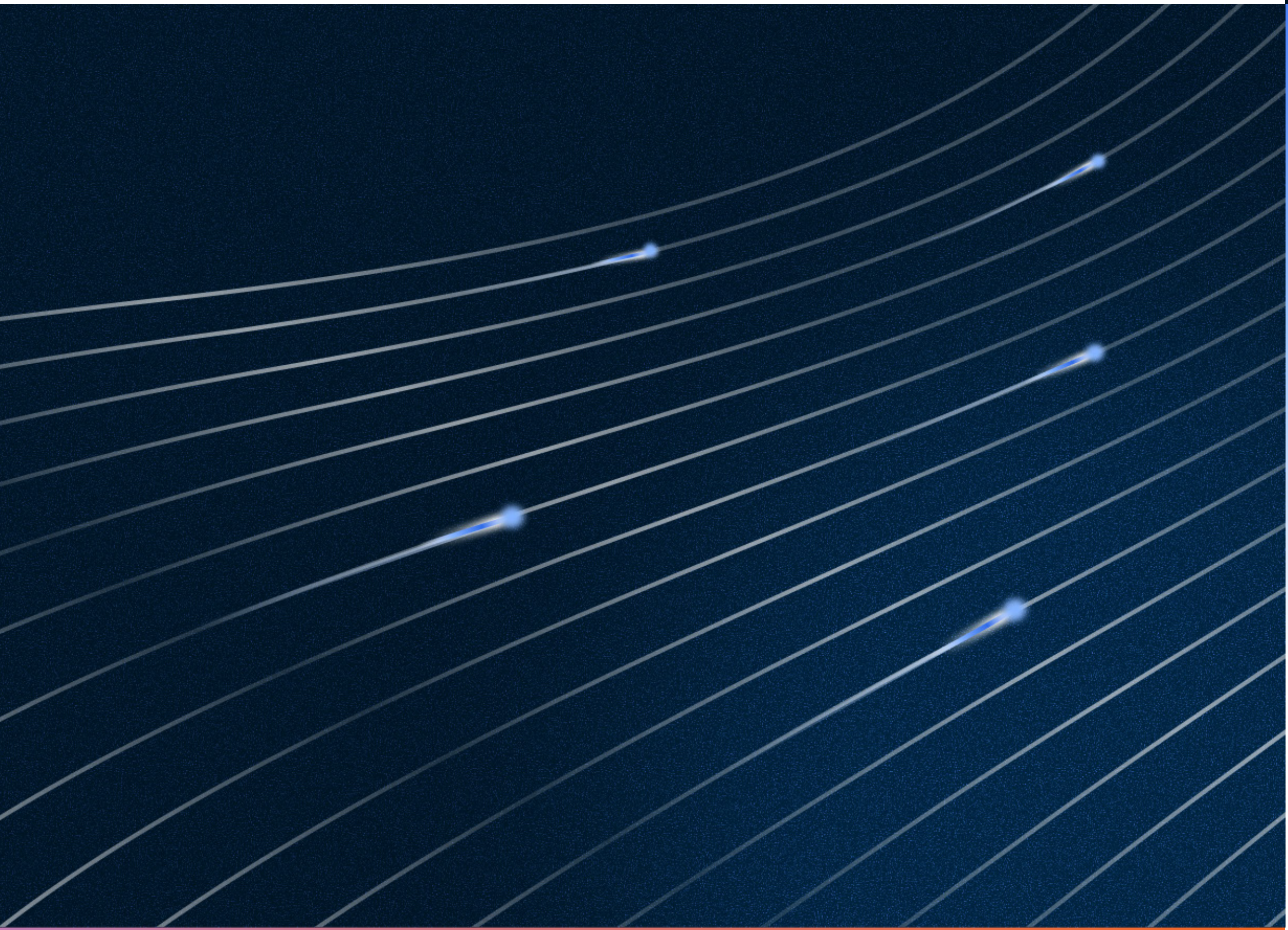


Table of contents

Executive summary

The job of a CDO has never been easy, with a median tenure of no more than 30 months.

Broadly speaking, successful CDOs accomplish three goals:

1. Ensure data access so that data can go from where it's produced to where it can be used.

2. Build innovative, valuable data products and use data to support every kind of downstream analytical and operational use.

3. Govern and manage data responsibly, ensuring that data usage is safe, repeatable, and accountable at all stages.

The ability of CDOs to execute these goals has long been hampered by unclear directives from company leadership, organizational immaturity, and a rapidly shifting technological and economic context.

The growing importance of AI has amplified these challenges. The stakes have risen — data is no longer primarily a matter of enabling analytics and decision support, but the essential prerequisite to intelligent automation across a business. Enabling your organization's use of AI means contending with entirely new categories of data products that impose additional, unique demands and requirements on your data infrastructure and data management practices.

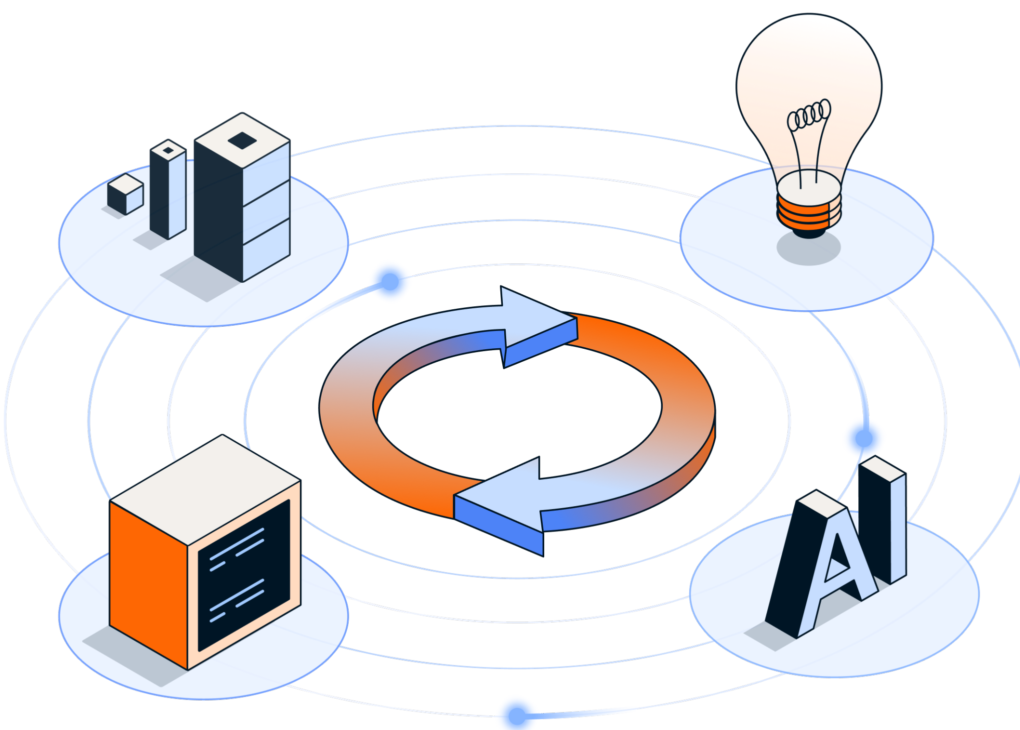
To solve the triple mandate of ensuring data access, building leading data products, and governing and managing data, CDOs need to:

1. Define and clarify their own mandate.

2. Translate organizational maturity from an aspiration into a series of concrete capabilities and benchmarks.

3. Build an interoperable, future-proof data foundation that can adapt to and absorb new AI and data use cases.

This guide provides the analytical clarity and concrete steps you will need to pursue these three solutions.



CHAPTER 1

Define the CDO’s mandate

Companies often appoint CDOs without a clear sense of what the role is supposed to accomplish, how success will be measured, how the rest of the organization will support the CDO, and what timetables are realistic for which outcomes. A lack of clarity can lead to unrealistic business expectations, poor KPI design, and limited buy-in (and therefore, support) from the rest of the organization.

A lack of clarity can also lead to overlapping or conflicting mandates with other C-suite offices, such as the CIO, CTO, CAIO, and CISO.

As previously mentioned, data operations must accomplish 3 key goals:

1. **Ensuring data access** — data must go from where it’s produced to where it can be used.
2. **Building innovative, valuable data products** — data must support every kind of analytical and operational use downstream, including AI.
3. **Govern and manage data responsibly** — usage of data must be safe, repeatable, and accountable.

These 3 key goals correspond with the following OKRs and their (non-exhaustive list of) KPIs. The following should give you a baseline for assigning ownership:

Goal	Objective	Key result
Ensuring data access	Readiness	AI-ready data coverage, including % of data assets and domains with clear • Owners • Definitions • SLAs • Quality thresholds • Lineage • Metadata • Access policies.
	Velocity	Time to value from ideation, including: • Time to onboard a new data source • Time to launch a new data product • Time from AI prototype to production
Building innovative, valuable data products	Value	Business value from AI products in production, including: • Revenue influenced • Hours saved • Churn reduction
	Adoption	Usage of AI products: • Monthly active users • Workflow penetration • Self-service query volume • Executive adoption • Satisfaction and trust scores
Govern and manage data responsibly	Trust	Governance and risk management: • Program compliance • Audit pass rates • Data lineage coverage • Data exposure incident rates

Any CDO with a properly defined role will own either all or a subset of this list. Ultimately, however ownership of these business goals, OKRs, and KPIs is divided, the key is to use data to enable the creation of business value through better decisions, operational efficiencies, automation, and new products. **The key challenge CDOs must solve, in part or in whole, is building durable data capabilities in a continuously evolving technology environment.**

As JoAnn Stonier, former CDO of Mastercard whose tenure lasted more than 5 years, once said:

"My goal as the data officer is to enable business. And I say this all the time, and people think I'm crazy, but you need to engage the business today and tomorrow's business. And that's a really tall order. And this is where I think data officers can be super strategic, but it means they have to understand their business and not only where it is today, but where it's going."

CHAPTER 2

What you will need

The capabilities necessary to enable a business with data range from higher-level architectural considerations to the specific capabilities of individual technologies, tools, and platforms. The key question to ask is not "which data capability should we build first?", nor is it appropriate to treat the acquisition of these capabilities as a one-time linear progression based on how data flows through a system (i.e., ingestion, quality control, governance, and products). It is equally impractical to attempt building all of your data infrastructure at once, not least because business needs, regulations, and technological capabilities are constantly changing.

Instead, focus on building a minimum viable product for delivering the highest possible reusable business value, then iteratively expand the scope and scale of your data operations. Reusability is key, as your capabilities must serve multiple teams and use cases, not just one project.

In short, you should prioritize capabilities based on the following criteria:

1. **Business value:** Which capabilities will increase revenue, productivity, or efficiency? Which can reduce risks or lead to novel products and services for customers?

2. **Reuse potential:** Which capabilities can be used by multiple teams and use cases, not just a single project?

3. **Risk reduction:** Which capabilities reduce risks associated with quality, privacy, regulations, security, and analytics or AI misuse?

4. **Adaptability:** Which capabilities will make it easier to absorb new tools and data sources on an ongoing basis?

5. **Time to value:** Can you show value within a meaningful, relatively short-term business cycle?

A practical sequence

Designate a high-leverage use case to start. This can start as something basic and practical, such as basic reporting and analytics, and become more ambitious, such as conversational analytics that combines reporting and analytics with a natural language interface.

A use case gives you a concrete, specific end goal while eliminating the temptation to do everything at once. Once you have proven value, you can expand the use case to adjacent domains (e.g., from revenue to product usage to customer health to sales productivity, etc.) and personas (e.g., from executives to managers, team leads, and ICs), as well as growing levels of technological sophistication (e.g., from dashboards to conversational analytics to an always-on assistant).

There aren't necessarily strict boundaries between each of the steps listed below, but they should give you a rough sense of what the basic "loop" for iterating and expanding your data infrastructure and use cases should look like.

1 Lay the groundwork for knowing and controlling data

Before anything else, you need basic visibility and control. Your organization needs to be able to see, govern, and coordinate all of its data assets to prevent confusion and sprawl. As AI grows in importance, issues such as metadata and lineage will only become more important as your organization faces the need to reconstruct (and prevent) hallucinations and other erroneous outputs.

The following capabilities will remain applicable as you handle or move data, and choose tools, platforms, and technologies.

- **Data cataloging:** You need to know what data assets you have, where each model lives, who owns it, and how it is used. This capability will grow in importance as data moves through your infrastructure and is transformed for context engineering or any number of data models.
- **Ownership:** Someone must assume accountability and ownership for each critical data domain.
- **Access control:** You need safe, governed access that can scale without relying on ad hoc approvals.
- **Governance policies:** Set explicit rules for privacy, security, retention, and usage.
- **Metadata foundation:** Aside from knowing what data assets you have, you also need context, summary metrics, and administrative details about them.

As you add tools in subsequent steps, keep these capabilities in mind. Consider them core, necessary features.

2 Start moving and modeling data

Once you have laid the groundwork for visibility and ownership, you will need several key capabilities concerning moving, integrating, modeling, and observing data.

- **Pipelines for data movement and ingestion:** You need to be able to centralize operational systems, SaaS applications, databases, event streams, files, and third-party data.
- **Data transformation, modeling, and context engineering:** Once raw data is centralized, you need the capability to collaboratively and systematically turn it into usable assets for analytics, operations, automation, and AI.
- **Orchestration, lineage, and observability:** As records are combined from multiple sources and transformed in multiple stages, you need the ability to organize and sequence the stages in ways that respect dependencies. It also goes without saying that you need to be able to audit the process, reconstruct failures when they happen, and correct them accordingly.
- **Semantics:** You need to systematically translate between data models and the real-world concepts they correspond to.

Fivetran and dbt Labs are strong advocates of the idea of the Open Data Infrastructure — an architecture characterized by scale, optionality, and interoperability. The keys to an Open Data Infrastructure are open standards, interoperable storage, decoupled compute, and unified context. By storing data once in open formats, data becomes portable, governable, and accessible across tools, engines, and AI systems. This combination of centralization and portability reduces duplication and lock-in while preserving control, flexibility, reliability, and long-term optionality.

Your choice of destination, in particular, will be pivotal. To complement the capabilities listed above, you will need the following:

- **A data lake using open table formats:** A data lake as a central data repository leverages the strengths of both traditional data warehouses, such as structure and governance, with the interoperability and scale traditionally associated with data lakes. Open table formats, such as Apache Iceberg and Delta Lake, are a layer of abstraction that wraps around data files like Parquet, organizing them into a database-like structure and providing capabilities such as ACID transactions, upserts and deletes, and schema evolution. However, ensuring that data is populated in the data lake in open table formats is non-trivial. You should strongly consider a managed service that ensures data is ingested in the appropriate format.

- **Compute engines that are ideal for your use case:** The interoperability of a data lake is predicated on the decoupling of storage and compute. In the past, your data team had no choice but to use the data warehouse's native compute engine. Now, table formats and metadata tell compute engines how to treat a collection of files as a table. Table formats track snapshots, manage schema changes, support concurrent reads and writes, and give transaction guarantees. The upshot is that you can choose any configuration and number of compute engines for any specific use case.

The modern data lake, which stores data in both open data formats and unstructured media files, is fundamentally a highly **reusable** platform that enables a single platform to support multiple data use cases at scale and cost-effectively, rather than duplicating data across separate silos. This ensures a single, consistent source of truth for all analytical and operational uses of data. As your data use cases grow in number and complexity, the ability to bring any tool to the data becomes a practical competitive advantage.

The alternative is multiplying infrastructure costs, synchronization problems, compliance and governance issues, and, most importantly, additional human bottlenecks.

As Andy Hill, whose tenure as CDO of Unilever lasted 7 years, remarked:

"Our advanced data platform combines our global, functional, local market data all together so we have one version of the truth and scale our analytics products efficiently. Having all of our data integrated allows us to improve the user experience, the speed and quality of insights, and reduce the total cost of data ownership."

3 Turn your data into products

Once data can be moved, observed, and governed, the next step is to make it consumable. Data products include dashboards and reports, data activations such as embedded analytics and automated workflows, predictive models, and AI agents.

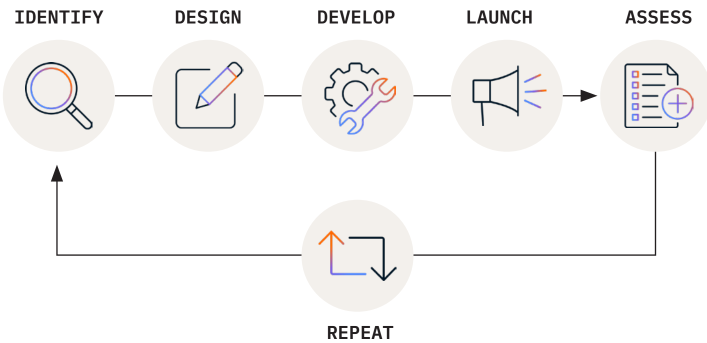
You must not only layer the appropriate tools on top of your data models but also incorporate **product thinking**.

- **Data as a product:** Package your data assets around business domains and use cases, with owners, SLAs, documentation, quality expectations, and consumers.

- **Lifecycle management:** Retire stale assets, version important ones, and track adoption and value.

- **Measuring success:** Measure usage, reuse, quality, latency, cost, and business impact.

The basic workflow for product thinking looks like so:



Identify

- Understand users
- Gather requirements

Design

- Define scope
- Manage expectations

Develop

- Rapidly prototype the product
- Ensure a repeatable process for continued development

Launch

- Market and roll out the product
- Train users as needed
- Drive adoption, including through self-service whenever possible

Assess

- Evaluate against expectations and KPIs

Turning data into products shifts the CDO's visible contributions from pure data engineering to clear productivity and business outcomes. Product thinking, more generally, allows you to systematize your pursuit of AI readiness and digital transformation into discrete, achievable goals.

CHAPTER 3

Preparing for AI and a changing future

There is no fixed end state of "being data-driven" or "AI-ready." Business models, regulations, AI capabilities, customer needs, and competitive landscapes will always change, which means that targets will always be moving. The Open Data Infrastructure is meant specifically to future-proof your data infrastructure, enabling you to combine commodity data storage with any sets of downstream tools as needed.

More generally, the very nature of AI poses both opportunities and challenges. As the focus of your analytics shifts from human decision support to automation through agentic AI, the demands on your data team and its infrastructure will change.

The most critical qualitative difference between data for a primarily human audience and data for a primarily machine audience is the continuous nature of automation. Humans work intermittently, periodically thinking through pivotal decisions before acting, and have office (and waking) hours. By contrast, AI agents operate continuously and can, in principle, perform work 24/7.

A corollary to this difference between data for humans and AI is that humans are equipped with a combination of practical know-how, judgment, and tribal knowledge, absorbed through learning, experience, and social interaction, that is not available to AI. Agents are far less sample-efficient than humans and must be explicitly provided with all necessary context, including relevant data, definitions, policies, history, and permissions. The importance of this context behooves data teams to centralize and unify it, ensuring that every human, application, and AI agent can work from the same trusted foundation and a single source of truth.

Moreover, current AI models lack the specific agentic training and general reasoning required to replace all human judgment. For the foreseeable future, humans will remain in the loop, but increasingly in a supervisory capacity, checking and acting on outputs created by AI systems. More than substituting for human labor, it is far likelier that AI will move human workers, including analysts, engineers, and knowledge workers of all kinds, up one or two layers of abstraction, coordinating certain streams of work rather than directly executing them.

For the foreseeable future, practical implementations of AI will be far more likely to leverage the power of foundation models by major providers (e.g., OpenAI, Anthropic, Google DeepMind, etc.) in conjunction with proprietary data. Neck-and-neck competition between major AI providers means that the foundation models themselves are largely commoditized, and what differentiates a high-performing implementation of AI will be the additional context it can access rather than the out-of-the-box capabilities of its foundation model.

As Teresa Heitsenrether, Chief Data and Analytics Officer at JP Morgan Chase, noted:

"AI models will emerge as a set number of models that are going to have a lot of similar capabilities. The differentiating factor is the data."

This introduces an ongoing governance challenge in its own right, as your organization will need to keep careful track of AI deployment patterns and their interactions with your organization's data.

How to think clearly about AI

You will not only need the reusable infrastructure and interoperability described earlier, but also a clear perspective on how your organization can use AI to enhance productivity or incorporate it into novel products.

Present-day AI has a "jagged" ability profile due to how it's designed and trained. The table below summarizes what AI is and isn't good at:

AI strengths	AI weaknesses
Pattern recognition and completion for text and code, including: • Drafting • Editing • Summarizing • Translation • Coding	Discerning truth from plausibility, especially when facts are obscure or highly specific, such as: • Citations • Recent events • Legal, medical, and other sensitive details
Ideation and brainstorming, especially when breadth is required, and the cost of a bad suggestion is low	Knowing when not to answer (e.g., when the premise of a request may be wrong or inappropriate, or there is no valid solution)
Reasoning through problems with clear, specified premises and constraints	Open-ended problems that require discernment, causal reasoning from limited evidence, and long-horizon planning and (especially) execution
Explaining and tutoring	Adversarial interactions

More generally, AI still faces an array of very real technical, business, and broader institutional limitations. As you determine practical use cases for AI, you must ask yourself the following sets of questions:

- **Technical feasibility:** Is this task or workflow something AI can actually perform well enough in real-world (not laboratory) conditions, given both the available data for training/augmentation as well as the characteristics of its architecture?
- **Economic characteristics:** Do the unit economics, task separability, and workflow integration make sense? Can an AI actually produce outputs faster, better, and cheaper than a human? This final question is especially nontrivial, given the potential for skyrocketing token costs.
- **Legitimacy:** How legally, socially, and politically acceptable is it for an AI to perform these tasks? Can the appropriate stakeholders be convinced? How can accountability and ownership be assigned?

These questions will help ground your AI projects and your communication of AI's capabilities, and also prevent the magical thinking and irrational exuberance (or despair) that frequently inflect discussions of AI.

CONCLUSION

A plan for continuing innovation and data-driven value

By clearly defining the CDO's role, cyclically and iteratively acquiring capabilities, and keeping a clear head about AI, you will be well-situated to outlast the CDO's nominal 30-month tenure. The Open Data Infrastructure, with automated data integration and context engineering as its keystones, will future-proof your data operations and ensure your organization's ability to adapt to continuing economic and technological changes. By combining the right data foundation with the right plan, you can ensure that your organization is at the bleeding edge of every development.

As the importance of AI has grown, the CDO's mandate has changed from managing data as an enterprise asset to making data usable as the foundation for intelligent, automated action. Before AI, the CDO's mandate centered on quality, governance, reporting, compliance, and analytics enablement; today, those responsibilities still matter, but are now table stakes for a larger mission: ensuring that data can safely and reliably power models, agents, automation, personalization, decision systems, and new products. That means the CDO must define the enterprise's data-and-AI mandate, build interoperable infrastructure for unknown future use cases, turn data into reusable products, embed governance into the data lifecycle, and help the C-suite understand that AI transformation is not primarily a model-selection problem, but a data readiness, trust, ownership, and operating-model problem.

Fivetran + dbt Labs deliver the data infrastructure layer that makes agents trustworthy — from the moment data moves, through every transformation, to the context an agent reasons from.

The **Fivetran** platform moves, manages, and transforms data from every system a business runs on into a secure, reliable foundation engineered to evolve, with the flexibility to work across clouds, engines, and tools. With Fivetran, analytics, operations, and AI run on data you trust and control. Thousands of organizations worldwide, including OpenAI, LVMH, Pfizer, and Verizon, rely on Fivetran to turn data into a competitive advantage.

Learn more at **Fivetran.com**, or follow Fivetran on **LinkedIn**.

Since 2016, **dbt Labs** has been on a mission to help data practitioners create and disseminate organizational knowledge. dbt is the standard for AI-ready structured data. Powered by the dbt Fusion engine, it unlocks the performance, context, and trust that organizations need to scale analytics in the era of AI. Globally, more than 100,000 data teams use dbt, including those at Siemens, Roche, and Condé Nast.

Learn more at **getdbt.com**, and follow dbt Labs on **LinkedIn**, **X**, **Instagram**, and **YouTube**.