Harbr_

CDO SURVIVAL GUIDE

Embrace complexity, accelerate value

CONTENTS

Introduction

Chief Data Officers are ultimately judged on the business outcomes they deliver. This is true whether they are addressing quality issues that generate regulatory risk, directly monetizing data to deliver new revenue, or any other business priority. And with the shortest average tenure of any role in the C-suite,[1] it's crucial for CDOs to understand and put into action concrete measures that deliver value to the business.

How can CDOs actually deliver against this fairly open-ended objective? What are the general areas worth prioritizing regardless of how value will be realized? How much time, money, and effort should be spent on different areas in pursuit of delivering value? Where should a CDO focus if they are to survive in their role?

The answers to these questions are highly dependent on the context of the environment in which the CDO is operating. Consider the impact that any of the following would have on the needs and priorities of a CDO:

1. Technology: The tools and systems used to collect, store, process, and analyze data.

2. Legacy: The existing data assets and historical data practices within the organization — and what this means for quality, accessibility, and compliance.

3. Regulation: Understanding relevant regulations, implementing necessary controls, and overseeing data privacy practices to mitigate legal risks.

4. Corporate Structure: How does this impact data governance and ownership?

5. Competency: The skills and expertise of the workforce in handling and deriving insights from data.

6. Culture: Attitudes, beliefs, and practices related to data.

The reality is that every CDO will be operating in a unique environment. The larger and more established the organization and industry, the more of these factors are likely to be in play.

While this list may provide a useful framework for understanding the context of a given environment, it provides little support for making the optimal decisions to deliver value. This white paper provides a view on what are likely to be the common truths for many CDOs, and practical guidance on how to prioritize capabilities that are strategically important while also addressing tactical needs.

This will be more relevant for larger, more complex environments where several of the above items will be a factor. As a result, those environments are likely to be the most challenging for delivering value sustainably and at scale. Arguably, these organizations also have the most to gain from getting it right. They tend to have very large data footprints, extensive commercial ecosystems, and expansive data investments that need to demonstrate a financial return.

[1] Brian Eastwood, "Chief data officers don't stay in their roles long. Here's why." MIT Sloan

Context is king

CDOs will already understand the nuances of those six areas in the context of their business. So rather than delve into them in detail, let's instead acknowledge that any of these in isolation generate complexity through their diversity.

CDOs must contend with diversity of the data itself, the users they need to serve, the technologies involved, and the organizational boundaries they need to cross:

Data: Quality, timeliness, format, structure, location, update frequency, governance, etc.

Users: Data scientists, analysts, engineers, domain experts, business leaders, data product managers, legal, compliance, etc.

Technology: Databases, source systems, file storage, cloud storage, data lakes, data warehouses, etc.

Boundaries: Teams, departments, legal entities, jurisdictions, partners, customers, suppliers, etc.

Much of the focus for solving complexity has been placed on technology. The reality is that no single technology can service the diversity of data and users at play or effectively transcend organizational boundaries. Perversely, efforts to remove complexity have resulted in further compounding the issue. The primary example of this has been the push towards data centralization. This approach has been criticized for offering a solution aligned to — and dependent on — a specific technology provider. It's also incredibly expensive.

More recently, federated and decentralized data architectures like data fabric and data mesh have gained support in trying to address complexity rather than remove it. However, these approaches have been criticized for lacking a tangible set of recommendations, leaving much to the would-be implementer to navigate and solve. Additionally, much of the technology that would be required to make that a reality at scale, and across complex environments, is nascent and continues to evolve. Consequently, most organizations already have a hybrid architecture and will continue to have one regardless of their current preference.

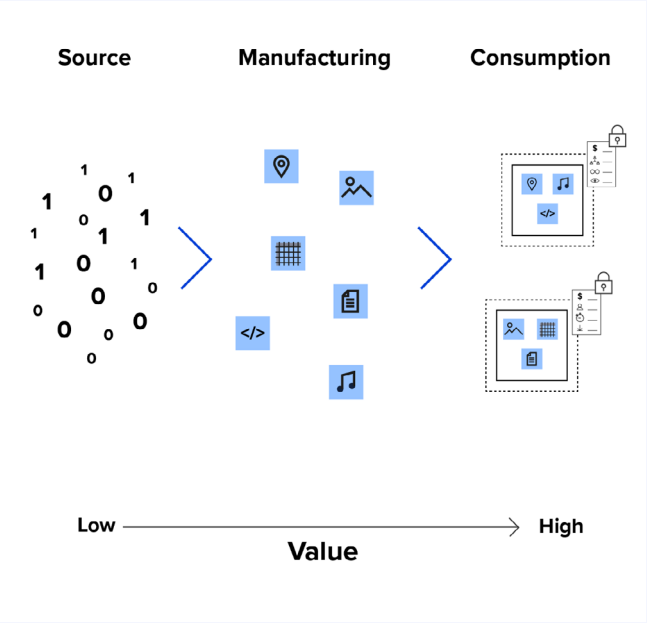
For full disclosure, Harbr is a technology that supports centralized and decentralized data architectures. This paper draws on our experience of supporting hybrid data architectures across some of the world's largest data ecosystems.

Let's start by exploring the challenges created by the diversity of data, users, technology, and organizational boundaries and considering some practical solutions.

Data

While it is fairly obvious to state that data is diverse, this relates to more than just format and structure. For a CDO, diversity in the utility and value of data is arguably more important to understand, as it should dictate where time and money are spent.

A useful framework for understanding this is to think in terms of a data value chain that spans raw data, data assets, and data products.



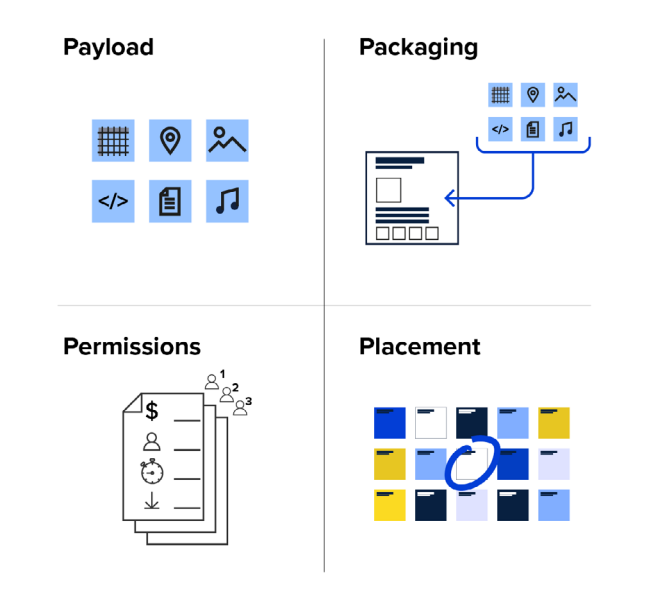
Data value chain: *Source systems (low value) to the consumption layer (high value)*

Raw data is data in its natural state, as generated at source. Often the data is a by-product of a source system and requires some level of 'manufacturing' in order to make it useful for a given use case. As a result, the value of raw data is typically low.

Data assets are digital objects that have been manufactured from raw data for a given use case. This could include tables, notebooks, SQL queries, PDFs, visualizations, etc. Data assets have the potential to deliver value, but this is typically limited by the lack of a scalable consumer experience (discovery, evaluation, usage, etc.). Data assets therefore have some value but this value is unlikely to have been realized to its full potential.

Data products are any combination of data assets, subject to active product management, that deliver a defined value proposition to a target market. The purpose of productizing data is to optimize value realization by enabling consumption at scale while minimizing cost. Consequently, data products are likely to be highly valuable.

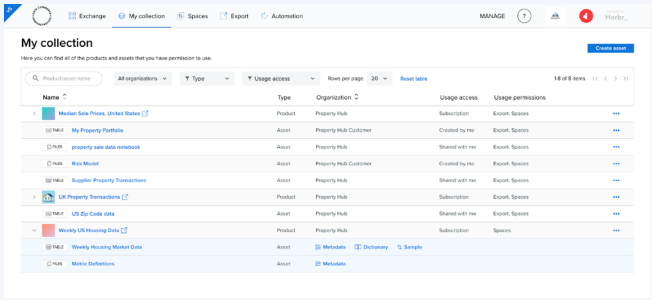
To optimize value realization, products will include packaging and permissioning, and will be optimally placed for the target market. This supports self-service experiences for discovering, evaluating, and using the product, which reduces cost. Crucially, this also supports consistent risk management, allowing for the product to be consumed by the widest range of consumers and across organizational boundaries.



4 Ps of data products: *Payload, packaging, permissions, and placement*

If the full value of a data asset can be obtained at a very small scale and without the need to cross boundaries, then its value can be optimized without the added burden of productization. Not every data asset can, or should, become a data product. Below is an example of an implementation of these concepts:

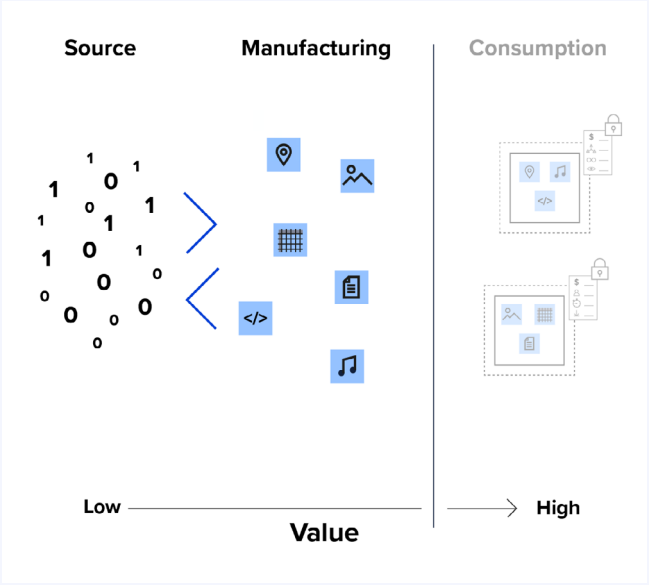
- There are individual assets that are both tables and files. The files include a model and a notebook.
- Two of the assets have been created by the user, and two have been created by someone else and shared with the user.
- The data product contains multiple assets.
- Access to the product (and the assets it contains) is via a subscription with self-enforcing rules to manage risk and value transfer.



Data assets and products: *Products contain assets such as tables and files on the Harbr platform*

This framework is helpful for considering where to focus time and effort. Clearly, the greater the number of products, the greater the value that is likely to have been realized. But products require assets that have been manufactured, and manufacturing requires raw data (or other data assets) as an input. Consequently, all of these data types must exist to some degree. The challenge for a CDO is to work out the right balance.

Unfortunately, many organizations get stuck in a manufacturing loop, with significant time and money spent creating data assets without a commensurate investment in value realization at scale. The normal balance is an overwhelming amount of raw data, a lot of data manufacturing, and a disproportionately small amount of value realization.



The manufacturing loop: *Effort is concentrated in low and medium value activities*

A lack of value realization capability can also create a vicious circle where data manufacturing becomes increasingly disconnected from the end-users and value propositions. The lack of interaction and feedback leads to increasing amounts of time manufacturing and maintaining relatively low-value assets, leading to a poor business case to invest in the capability to change that. A key defense against that is to focus on and measure 'time to access' for data consumers, e.g. "How long does it take the average data consumer to get the data they need?"

> Data assets can deliver value that is typically limited by the lack of a scalable consumer experience for discovery, evaluation, usage, collaboration, and delivery.

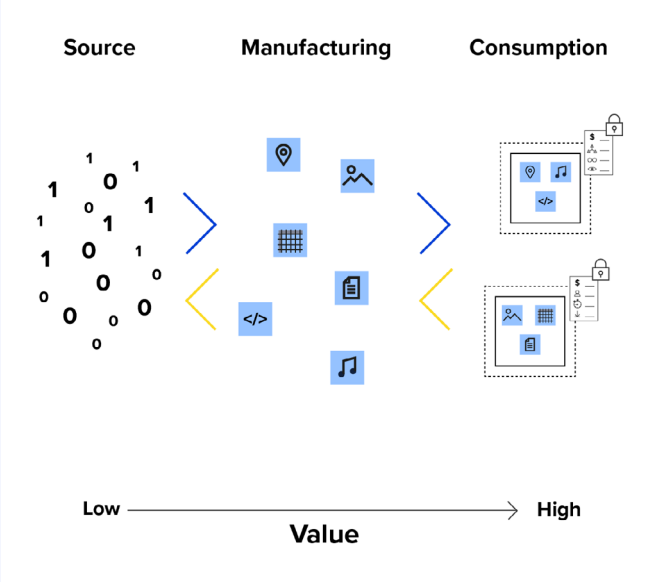
Data access

In most organizations data access is slow, painful, and expensive, often in spite of large data investments. Access is a critical step in realizing value from data, so if that is inefficient, everything stemming from it will be too.

Prioritizing data access starts a virtuous circle that empowers data consumers to deliver value, generates feedback for those manufacturing the data, and neatly dovetails with good product management practices where value needs to be optimized.

This feedback will provide detailed, real-world requirements about quality, timeliness, format, structure, location, update frequency, governance, and other dimensions — all of the aspects that need to be factored into data manufacturing.

Once you've got that feedback from multiple use cases, you can prioritize manufacturing only the assets that make economic sense and only to the extent required to deliver the target value — a cost-benefit analysis that wouldn't be achievable without this iterative process.



Access-driven feedback loop: *With added focus on data access, a virtuous feedback loop allows needs and insights from data consumers to inform activities at the source and manufacturing layers.*

Formats and structures

Away from this framework, it remains important to consider the diverse types and formats of data that can exist in any large organization:

CSV, JSON, Avro, ORC, Hierarchical Data Format, XML, NetCDF, FITS, Apache Parquet, text, binary, STL, GRIB, etc.

This bewildering array of formats and structures exists for a reason. Different use cases, different technologies, and different users require different formats to support their needs. One approach to this is to try and reduce diversity by re-engineering data into certain formats and structures to make it easier to use and get value from.

While in many cases this may be true, aiming to complete this task in advance of value realization risks ending up in the manufacturing loop mentioned earlier. Unless you work in a small environment with little to no legacy, the standardization of data will be a Sisyphean task and a new, better format is always just around the corner. The opportunity cost will therefore be the ability to realize any value whatsoever. Instead, undertaking this work iteratively and in line with a deep understanding of the target use cases is a more pragmatic approach that's likely to yield significantly more value.

Most large organizations have significant data diversity, both in terms of what data delivers value and the formats and structures of the available data. CDOs should not prioritize reducing diversity; instead, they should develop strategies for delivering value regardless of it.

Data diversity: practical solutions

Do:

- Understand the level of data diversity that exists and why.
- Determine how much data is valuable and how much is ready for scale consumption.
- If data is not valuable, limit the time and cost of storing and securing it.
- Identify the issues preventing rapid data access, and therefore, value realization.
- Implement processes and technologies to increase value realization.
- Ensure that data manufacturing is commensurate with value realization.

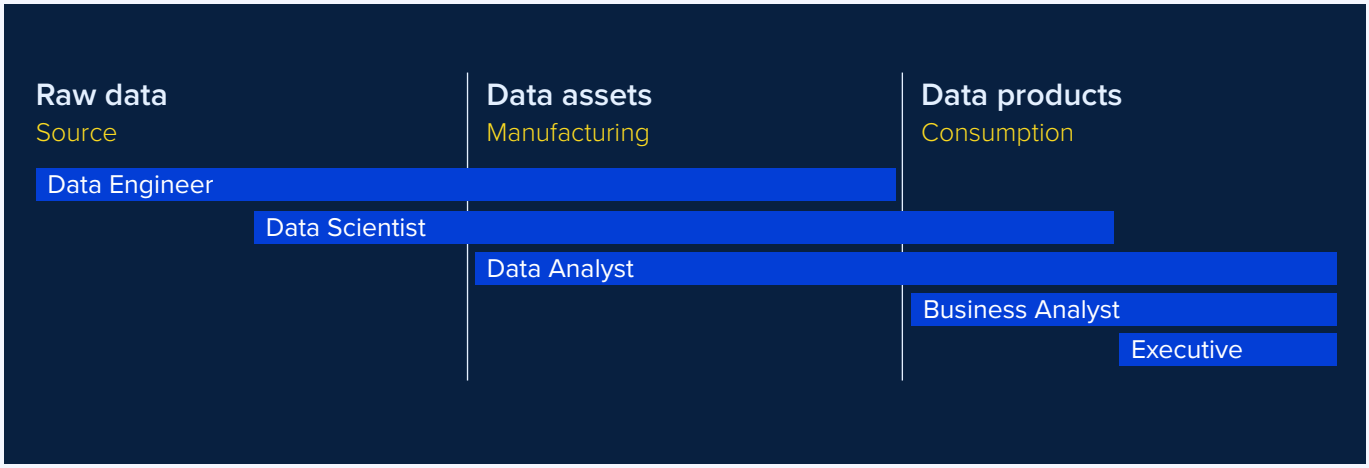
Don't:

- Believe that raw data is valuable.
- Manufacture data by default without a clear use case.
- Prioritize data manufacturing ahead of value realization.
- Unnecessarily manufacture data as this adds to the complexity rather than reduces it.

Data products are a mechanism for optimizing value realization by enabling consumption at scale, and across boundaries.

Users

Your organization is no doubt already serving a diverse set of users. Their different skills and priorities will necessitate varying consumption experiences across your data assets and products. Here's a view of the key groups of users and their typical participation across the data value chain.



Users on the data value chain: *Each role spans different parts of the data value chain*

The technologies, interfaces, and data assets required by these various stakeholders are very diverse. You can't teach everyone the same coding languages, and no technical interface will support each user.

Ideally, you will have the capability to broker access to a wide range of data assets for each group of data consumers. That should operate independently of the data manufacturing layer and data consumption tools, which are diverse and change rapidly.

Role	Data assets used
Data scientist	Notebook assets, code, file assets
Data engineer	Python, SQL
Data analyst	SQL, BI tooling
Business analyst	Visualizations, dashboards, spreadsheets
Business executive	Dashboards, PDF reports

Producers and consumers

A useful framework to examine your data users is to think in terms of whether, and when, they are producers or consumers.

Data producers include every individual involved in creating data assets or data products with a view to realizing value. Data consumers include every individual who accesses and uses data in any form (raw, assets, or products) including for the purpose of realizing value. The reality is that many data users in a typical organization are both consumers and producers.

This can be as part of the data value chain where they are consuming raw data or a data asset and refining it to move closer to realizing value. Alternatively, this could happen asymmetrically where their producing and consuming activities have no relationship whatsoever.

To avoid over-simplification, it's important to acknowledge that users do not only sit within the value chain, but their position is also affected by their activities related to producing and consuming. All data consumption activities, regardless of user type, will benefit from a dedicated capability that manages the interface between the producer and the consumer. Crucially, that capability must be able to serve the broad spectrum of needs for the relevant user types.

The capability for technical consumers involved in manufacturing raw data into data assets is typically ETL tooling and is very well-developed. The capability for technical and non-technical users to access and use data products has only recently been developed and proven at scale. A great example of this is Moody's DataHub, an award-winning data platform built on Harbr. You can learn more in the case study.

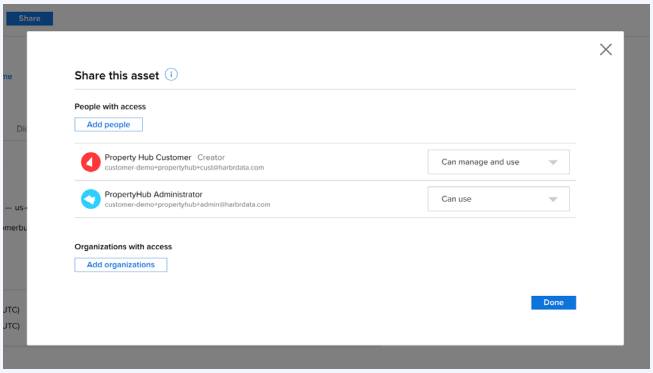
Risk and value

As previously mentioned, there are two important dynamics at play between producers and consumers:

Value transfer: To sustainably deliver success, the producer and consumer require a win-win scenario. The ability to adapt and optimize value transfer will support value realization. Whether or not there is a direct financial transaction involved, various mechanics, such as return on investment (ROI), budgets, and resource priorities will be at play and must be managed at the interface between consumers and producers when data assets and data products are involved.

Risk management: Similarly, enabling access and use of a valuable asset is likely to carry risk to both the producer and the consumer that must be managed. With data, this will typically be a diverse range of risks involved including security, legal, regulatory, ethical, and technical.

Here is a practical application of two risk frameworks — one for data assets and one for data products — that both provide a variable level of control. This variability not only allows for the right level of value transfer and risk management, but also differentiates between the level of control required for a data asset versus a data product. At any point, a data asset could be productized to benefit from the extra controls if required.



Sharing a data asset: *User sets usage rules when sharing a data asset with other users or organizations*

Data product subscription: *Subscription rules are enforced around lineage, usage, duration, and price*

To provide more detail, the subscription-based access control for the data product includes the ability to control the following dimensions:

Capability	Value transfer	Risk management
Terms & conditions	✓	✓
Number of users	✓	
Controlled, permissioned access		✓
Pricing	✓	
Control over storage, copies, movement of files, export rules, etc.	✓	✓
Revocable, timed access	✓	✓
Auditability around access and usage		✓

Once specified, these items should be enforced, ideally automatically, in the interface between producers and consumers - essentially a self-enforcing data contract that would ideally be readable across systems.

Data product management

As previously mentioned, data products require '*active product management*'. What does this mean in practice?

Clearly, in order to make and manage data products, you will require data product managers. Taking a product management approach to your data assets and products will help establish line of sight between producer and consumer, deepening the understanding of users and use cases.

The purpose of a data product - like any product - is to optimize value, which can be open-ended. The data product lifecycle is overseen by the data product manager, whose job is to develop a clear lifecycle for data products, including the planning, development, testing, deployment, and retirement phases. They must evaluate the viability of data products over time and consider sunsetting products that are no longer valuable. They should consider how products can be adapted to different use cases and make efficient use of common data assets such as tables, notebooks, and queries.

Data product managers regularly undertake the following activities as they build compelling data products:

- Seek to understand the available assets and any constraints attached to them e.g. technical, legal, cost, etc.
- Hypothesize the potential value of data products that could be built and rank them accordingly.
- Prototyping data products and collaborating with would-be consumers to understand and prioritize the specification, use cases, and value proposition.
- Building a minimum viable product (MVP) that addresses a specific business problem that needs to be solved — a defined value proposition.

Establishing a virtuous feedback loop between data producers and consumers will help you make more strategic decisions about technology, which we talk about in the next section.

Value transfer occurs when data is exchanged between users or organizations. Resource priorities between consumers and producers must be managed when data assets and data products are involved.

User diversity: practical solutions

Do:

- Understand where different data users are in the overall value chain.
- Understand where and when different users are consumers, producers or both.
- Prioritize the needs of different personas based on their contribution to value realization.
- Acknowledge that different types of data users will always exist and are required to realize value
- Use technologies to create a dedicated interface between producers and consumers.
- Invest in data product management capabilities to drive the value realization agenda.

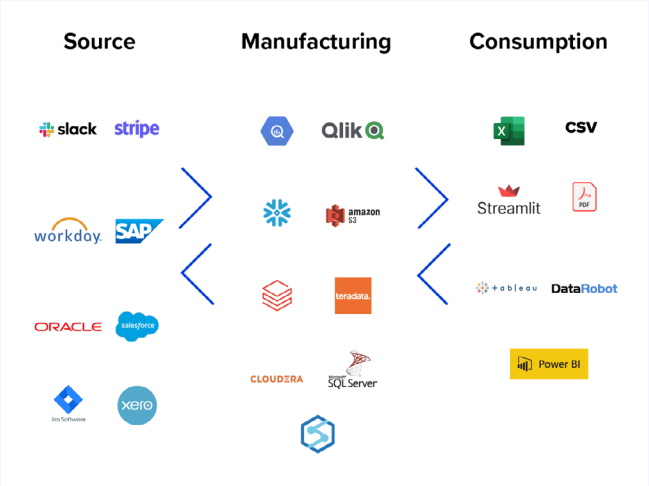
Don't:

- Make it a priority to train everyone to a high level of data competency.
- Force different users to interact with data that is not optimal for their skills or needs.

Technology

For a long time, the role of IT has been to rescue unit costs and ensure systems availability. This is typically achieved by minimizing the number of systems to reduce risk and realize economies of scale.

Despite this, most large organizations continue to have highly diverse IT estates — complexity remains the norm despite significant efforts to reduce it. Let's look at a simple framework for understanding the three layers of technology in the data stack:



Systems on the data value chain: *The data ecosystem broken out into source systems, the data manufacturing layer, and the consumer layer.*

Data consumers will need to gain access to data that originates from source systems in order to do their job. The more systems there are in between, the more challenging that becomes. And since scale implies diversity and complexity, extracting and using data from source systems is typically slow, hard, risky, and expensive.

Data manufacturing

Many data consumers won't typically access source systems, because they contain raw data that is not yet of any value. Some consumers, such as data engineers, will be comfortable with this but many won't so the norm is to have an intermediate 'manufacturing' layer.

The data manufacturing layer includes data warehouses, lakes, and a wide range of database types for specific use cases. This broadly maps to where data assets are created, stored, and managed, and where data assets are accessed by data consumers with technical skills.

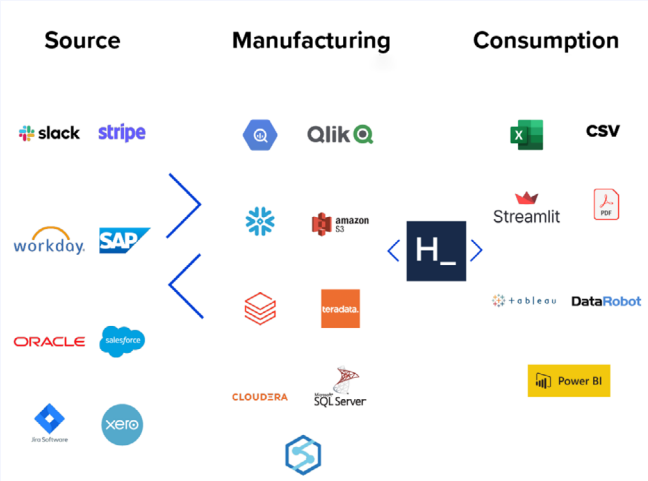
Data consumption

The data consumption layer includes tools used by data consumers, which are diverse and include everything from technical categories like AI/ML tooling, to Business Intelligence, spreadsheets, and even PDFs. Consumers typically have preferred tools and will be reluctant to change their technology preferences simply to access data.

Consequently, data producers will need to be able to meet the consumers where they are. Given the diversity of users and technologies at play, this can be a daunting task for any data producer and a significant hurdle to value realization.

It's therefore crucial to manage technology diversity by focusing on how to live with it, as opposed to trying to avoid it through centralization. While there are benefits to some level of centralization and standardization, this has continually failed to solve the problem, and in some cases has made it worse.

A solution to managing technological diversity at the intersection between data manufacturing and data consumption has emerged: self-service, on-demand environments where producers and consumers can access and use data assets and products regardless of where they're stored.

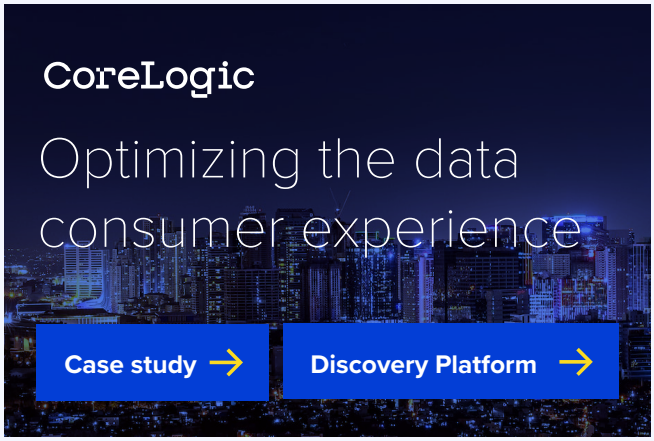


Harbr on the data value chain: *In this example, the Harbr platform acts as an interface between the data manufacturing and data consumption layers. Harbr allows the organization to control access to data on a subscription basis.*

Real-world example

A working example of this can be seen in the way CoreLogic, the market leader in property data, delivers data science services to its clients. CoreLogic's Discovery Platform, powered by Harbr, allows consumers to evaluate and trial data products, overlaying their own data to create custom outputs that deliver value. This takes place in secure cloud workspaces, meaning that each party can retain custody and control over their data assets. This technical capability has led to a deeper relationship between producers and consumers, resulting in higher-value data products.

Jonathan Gallo, Discovery Platform's Principal Product Manager, explains, "The Discovery Platform has redefined our client engagement model. Now we're deeply involved in our clients' journey from the very beginning — clearly understanding their goals and their challenges — continuing with them throughout every step of that process until they have something that they can implement as their solution."



Technology diversity: practical solutions

Do:

- Acknowledge that different technologies are performing different roles, for different data types, and different producers and consumers.
- Understand the most valuable use cases before designing architecture or investing in new manufacturing technologies.
- Continue to both aggregate data and leave it at source depending on what is best for the use case.
- Invest in building a dedicated layer to manage interactions between the manufacturing and consumer layers.
- Use the consumer layer to get feedback on what should be manufactured vs. consumed from source.

Don't:

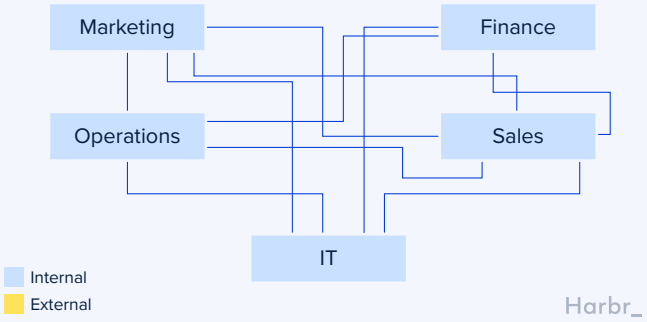
- Assume a single manufacturing technology will meet all your producer and consumer needs. None has to date...
- Be dogmatic about using a single architectural approach for your data stack.
- Invest heavily in data architecture at the expense of technologies that enable value realization.

Organizational boundaries

In addition to a diverse data, user and technology landscape, most CDOs will operate in an environment full of boundaries that introduce complexity. Both data consumers and producers will have to find ways to traverse those boundaries if they are to successfully realize value at any meaningful scale. This includes:

Internal

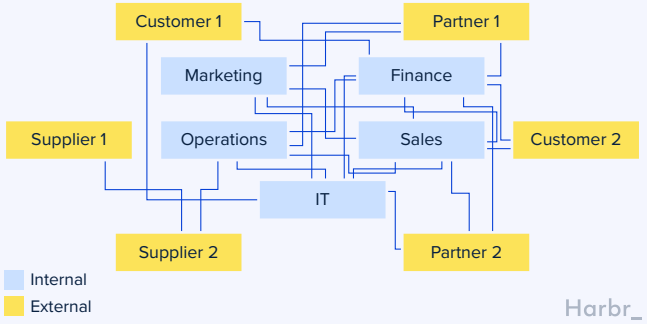
Teams, departments, corporate group: Working with 'internal' stakeholders who may have competing priorities and will need to consider cost and resource allocation. Friction here can make data access slow and a lack of supporting technology can make it difficult and expensive. There will often be large backlogs of data requests being serviced in an unscalable way.



Internal organizations share data with each other in an inconsistent, ad hoc manner.

External

Partners, customers, suppliers: External parties can be both providers and consumers of data. There generally has to be a joint value proposition for data access and use to occur. This can take many forms, including a direct financial transaction, a shared benefit, or a public good. Establishing and fulfilling the joint value proposition is often difficult and expensive. Engaging with external entities arguably carries more risk and certainly requires more robust risk management. The result is that external data sharing and access is typically slow and expensive in the absence of supporting processes and technologies to manage risk and value transfer.



Internal and external organizations in the data ecosystem.

National/International

Jurisdictions, laws, regulations: Crossing national boundaries can often introduce new risks and issues. Data may not be legally allowed to leave or enter a given jurisdiction for legal, reputational, or regulatory reasons. Understanding what data can and cannot traverse a national boundary and for what use cases, is a challenging and highly complex task involving many different stakeholders. It is also subject to change as laws and regulations change. As a result, any enabling technologies will need to provide flexibility and transparency around data storage, access, and use to effectively enable data-driven outcomes.

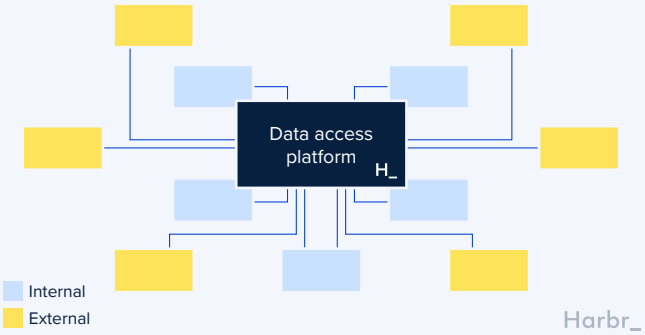
Boundaries: practical solutions

Do:

- Invest in a data sharing framework that can be embedded into access permissions.
- Invest in technology that can support both access and custody to provide flexibility around value transfer and risk management.
- Terms and conditions of access should be embedded as close to the data as possible and ideally enforced by technologies.

Don't:

- Ignore boundaries. They are a significant constraint and regularly change.
- Try to create a new legal structure or centralized team to solve it.



The data ecosystem is ideally connected by a platform for access, sharing, and collaboration on data.

Conclusion

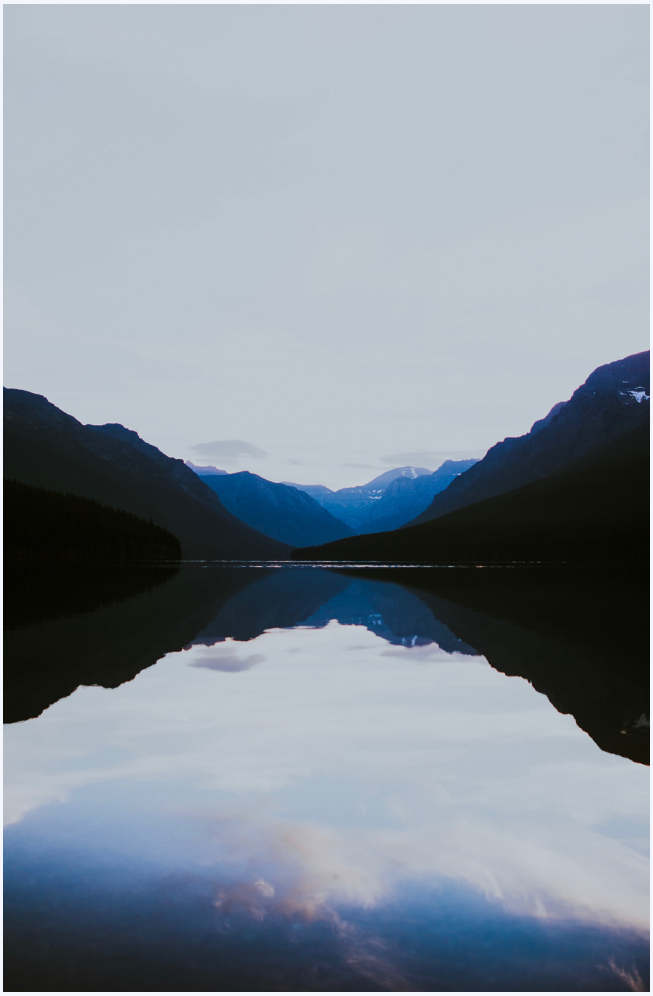
Chief Data Officers have the potential to drive enormous value for their organizations. The problem is, so much of that potential is out of their direct control. To buck the trend of unmet expectations and short tenures, CDOs must overcome a set of challenges around the complexity created by diversity — the diversity of the data, of the users they serve, the technologies they work with (not always out of choice), and the boundaries they need to traverse.

The data ecosystem at any large organization is highly complex, with considerations ranging from quality and timeliness to format, structure, and governance. Users come from diverse backgrounds, including data scientists, analysts, engineers, domain experts, business leaders, and data product managers, each with unique requirements and expectations. The technology stack is constantly evolving, encompassing databases, source systems, cloud storage, data lakes, and data warehouses, among others. Organizational boundaries are a complex web of teams, departments, legal entities, jurisdictions, partners, customers, and suppliers in a constant state of flux.

This guide seeks to provide pragmatic guidance for CDOs and data professionals on navigating the intricacies of data management within the context of these constraints. By exploring the challenges and solutions related to the diversity of data, users, technologies, and boundaries, we hope this helps you to quickly and sustainably realize value by focusing only on what matters most.

Learn more about realizing value from your data with Harbr's award-winning data platform.

Book a demo →



Appendix

Practical Solutions

Do:

- Understand where different data users are in the overall value chain.
- Understand where and when different users are consumers, producers, or both.
- Prioritize the needs of different personas based on their contribution to value realization.
- Acknowledge that different types of data users will always exist and are required to realize value
- Use technologies to create a dedicated interface between producers and consumers.
- Invest in data product management capabilities to drive the value realization agenda.
- Invest in a data sharing framework that can be embedded into access permissions.
- Invest in technology that can support both access and custody to provide flexibility around value transfer and risk management.
- Terms and conditions of access should be embedded as close to the data as possible and ideally enforced by technologies.
- Understand the level of data diversity that exists and why.
- Determine how much data is valuable and how much is ready for scale consumption.
- If data is not valuable, limit the time and cost of storing and securing it.
- Identify the issues preventing rapid data access, and therefore, value realization.
- Implement processes and technologies to increase value realization.
- Ensure that data manufacturing is commensurate with value realization.

Don't:

- Make it a priority to train everyone to a high level of data competency.
- Force different users to interact with data that is not optimal for their skills or needs.
- Ignore boundaries. They are a significant constraint and regularly change.
- Try to create a new legal structure or centralized team to solve it.
- Assume a single manufacturing technology will meet all your producer and consumer needs. None has to date...
- Be dogmatic about using a single architectural approach for your data stack.
- Invest heavily in data architecture at the expense of technologies that enable value realization.
- Believe that raw data is valuable.
- Manufacture data by default without a clear use case.
- Prioritize data manufacturing ahead of value realization.
- Unnecessarily manufacture data as this adds to the complexity rather than reduces it.

PUT YOUR DATA TO WORK

harbrdata.com

Harbr_