EBOOK

Understanding Data Debt:

A growing concern for data professionals



matillion

Contents

01 Technical debt vs. data debt

02 What contributes to data debt?

03 The main categories of data debt

04 Measuring data debt

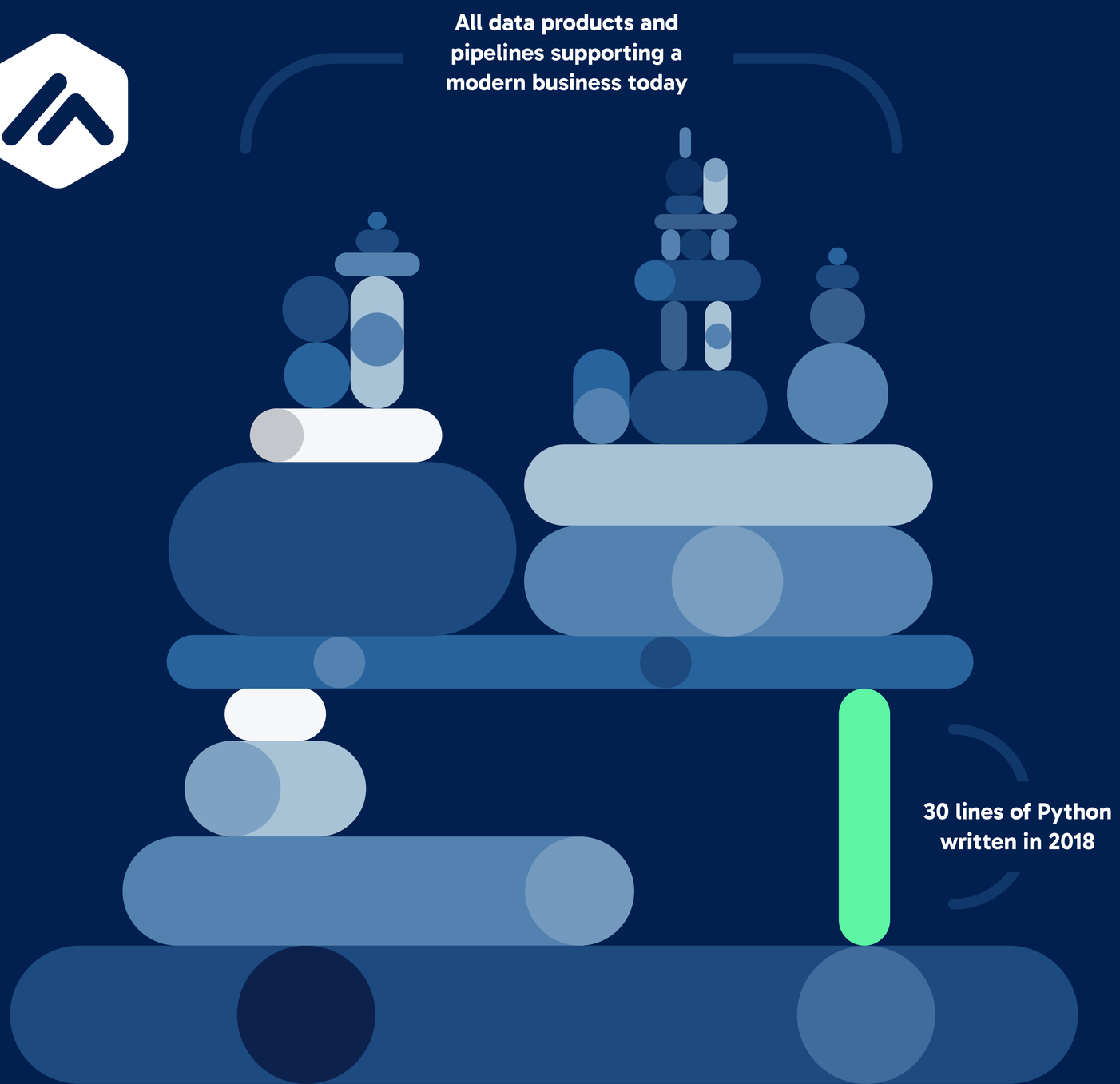
05 Examples of data debt

06 Business consequences of data debt

07 Managing data debt

08 Conclusion





Introduction

Data debt is the build-up of data-related problems over time, including quality, literacy, and security issues. Like technical debt, data debt represents the cost of avoiding or delaying investment in maintaining, updating, or managing data assets, leading to decreased efficiency, increased costs, and potential risks. The consequences of data debt are high: it creates a lack of reliable and repeatable data management processes and hampers productivity across teams.

This guide provides a comprehensive view of data debt: what it is, how it accumulates over time, real-world examples, and how you can take steps to remove data debt from your organization.



01

Technical debt vs. data debt

While technical debt and data debt can both impede organizational efficiency and effectiveness, their scopes are distinct.

Technical debt is primarily associated with code and software systems that require maintenance, updates, or replacements. Data debt, however, is broader, encompassing everything related to data assets. This includes data storage, processing, governance, quality, security, and even the organizational culture surrounding data.

Technical debt refers to the accumulated cost of suboptimal technical decisions made during software development or tools that were purchased. It encompasses issues such as outdated or poorly written code, lack of documentation, and the need for refactoring or reengineering portions of a product or service.

Data debt is the accumulation of issues related to data management, including data quality, governance, and security problems. It arises from inadequate data governance practices, inconsistent data management, and lack of a data-driven culture.

Neglecting data maintenance leads to inefficiencies and risk – similar to the consequences of ignoring technical debt. Data debt can result in poor decision-making, increased operational costs, and lost opportunities. The inefficiencies caused by data debt can manifest in various ways, including slow data processing, inaccurate analytics leading to bad decisions, and the high cost of maintaining legacy systems.

EXAMPLES OF...

Technical Debt

- Code that is not scalable or maintainable due to shortcuts taken during development
- Systems that require extensive updates or rewrites to incorporate new features or improvements
- Legacy software that no longer meets current business needs but is still in use because replacing it is costly and time-consuming

Data Debt

- Data silos result from fragmented data management across different departments
- Poor data quality due to lack of validation, cleansing, or governance
- Inconsistent data formats and standards that complicate data integration and analysis



02

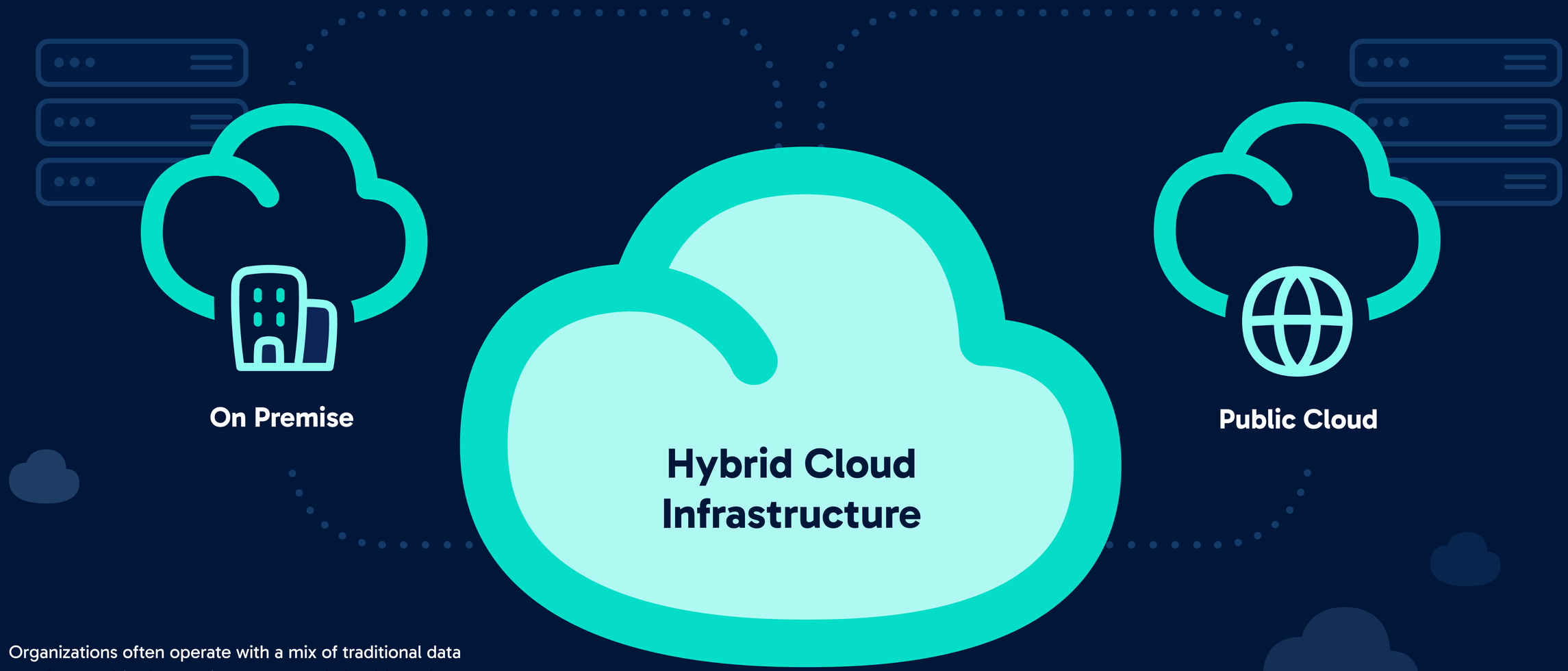
What contributes to data debt?

How data debt developed over time

Where and how organizations store data has changed dramatically over the past few decades. Beginning in the 1980s, data was predominantly managed within highly structured, conformed environments such as data warehouses. With the advent of big data technologies and the subsequent shift to data lakes in the early 2000s, data volume and variety of data expanded exponentially and became harder to control or govern.

Most recently, the migration to cloud platforms has further diversified the data ecosystem causing even more duplication of data and more challenges to data management, governance, and control. This evolution has significantly increased the complexity of data management.





Organizations often operate with a mix of traditional data warehouses, big data environments, and cloud-based systems. This hybrid data landscape, while offering numerous advantages (data mobility, security, and cost for example), also contributes to the accumulation of data debt.

Business factors that contribute to data debt

01

Accumulation of vast data over time and the rise of data silos

As organizations accumulate more data, the potential for creating silos increases. These silos often result from isolated data management practices across different departments or teams, leading to fragmented and inconsistent data.

02

Lack of data governance

Without robust data governance frameworks, organizations struggle with data inconsistencies and redundancies. Poor governance leads to data quality and security issues, making it difficult to trust and effectively use the data for decision-making.

03

Low data literacy and lack of a data-driven culture

When team members do not understand how to manage and use data properly, it results in poor data practices and the perpetuation of data issues. Low literacy also means data professionals could be using the wrong data sets or outdated data for analytics and AI or just keep duplicating the same data over and over because they just don't know what they can use.

04

Fragmented data management across multiple teams and architectures

Data management can consist of everything from ingestion tools, catalogue tools, data storage, and processing. Data management is often fragmented across different teams and architectures in large organizations. This fragmentation can lead to inconsistent data practices, making maintaining data quality and governance challenging.

05

Challenges from rapid organizational growth or mergers and acquisitions

Rapid growth or mergers and acquisitions can introduce new data sources and systems, increasing the complexity of data management. In these situations many times there will be duplicate systems or competing systems for data management causing tension and long painful migration projects often taking years to resolve. Integrating disparate data systems often results in data silos and inconsistencies, contributing to data debt.

03

The main categories of data debt



Technology surrounding the data: storage, processing, and architecture

Over time, outdated technologies and architectures can accumulate, leading to inefficiencies and increased costs. Over the last couple of decades, there have been so many different innovations in data storage, processing, and general data management including the Big Data and Hadoop era moving into the Cloud and now the AI era.



The data

Data quality, age or freshness of the data, governance, and security are critical components of data debt. Poor data quality and lack of governance can significantly impact an organization's ability to leverage data effectively.



Data Culture or Organizational Approaches

A data-driven culture is essential for effective data management. Organizations with low data literacy and poor data practices are more likely to accumulate data debt. Without a good data culture, all the negative data issues mentioned are guaranteed to happen in organizations of any size.



Shadow IT

Shadow IT refers to the use of information technology systems and solutions without explicit organizational approval. This often leads to the creation of data silos and inconsistent data practices, further contributing to data debt. Because the central data organizations are unaware of these pop-up IT organizations, they often purchase tools and data platforms to satisfy business objectives. This has led to duplication of efforts, possibly platforms that do not meet security guidelines, provide no data quality assurances (and many more issues).



04

Measuring data debt:

Steps to gauge severity, impact, and cost

01

Identify the severity of the data debt

This involves assessing various aspects such as efficiency, productivity, profitability, inaccurate decision-making, and missed opportunities due to data inaccessibility or slow report generation. Severe data debt can lead to substantial delays in generating new data products or reports, hampering business agility and decision-making.

Consider a scenario where a company's sales team relies on outdated and fragmented data pipelines to generate sales forecasts. The inefficiencies and inaccuracies in these pipelines lead to incorrect forecasts, resulting in poor strategic decisions and missed sales targets. This type of data debt is highly severe due to its direct impact on profitability and decision-making.

02

Evaluate the impact on your organization

Understanding how data debt affects your organization helps prioritize which debts to address first. Focus on areas where data debt poses security risks, leads to missed opportunities, or causes bad decisions due to poor data quality. Identifying low-hanging fruit—quick and easy fixes that can resolve multiple issues—can also be highly effective.

For example, if a core dataset used across multiple departments is outdated, updating this dataset can resolve numerous downstream issues, improving overall data quality and operational efficiency. Prioritizing such impactful debt can provide immediate and significant benefits.

03

Review the cost of addressing the debt

What are the resources needed to remove or address the data debt? This could include new platforms, systems, processes, or simply fixing ingestion or transformation code and pipelines. Sometimes, addressing data debt might involve decommissioning legacy systems that pose security risks or incur high maintenance costs.

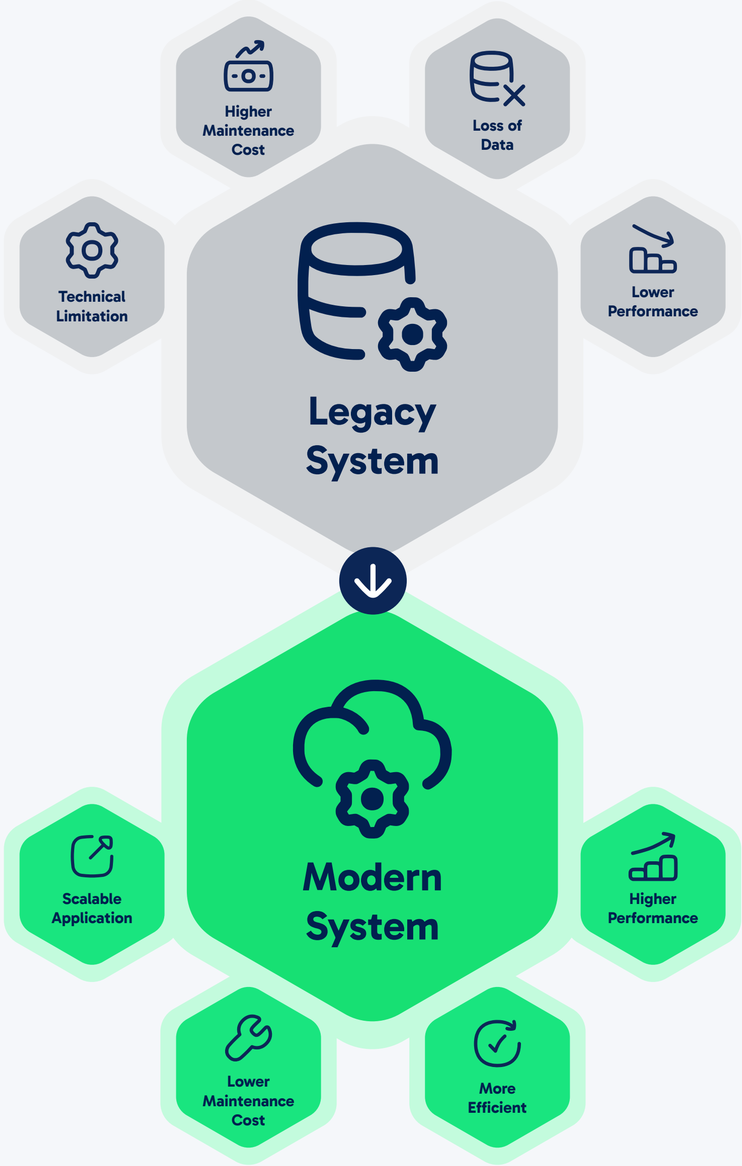
For instance, a company may have several legacy systems with overlapping data, leading to confusion and inefficiencies. Consolidating these systems into a single, modern data platform might involve significant upfront costs but will streamline operations and reduce long-term maintenance expenses.

04

Prioritize data debt

Once you have evaluated the severity, impact, and cost of data debt, prioritize which debts to address first. Sometimes, it is not just cost that will make one issue rise to the top; if something poses a risk to new emerging projects, like implementing GenAI applications, those should rise to the top. Start with debts that offer the most significant benefits or pose the greatest risks if left unaddressed.

A practical approach might involve categorizing data debts into high, medium, and low priority. High-priority debts would include those with severe impacts on security, decision-making, or profitability. In contrast, low-priority debts might involve less critical issues or those that require minimal resources to resolve.



05

Examples of data debt

Outdated and incorrect reports

Unused datasets and reports can lead to wasted resources and confusion. For example, an organization using Tableau or Power BI might have numerous outdated dashboards created over the years. These obsolete dashboards can clutter the workspace, making it difficult for users to find relevant reports, thereby reducing productivity and increasing frustration.

Regularly auditing and decommissioning unused dashboards can help maintain a clean, efficient reporting environment, ensuring that users only access current and relevant information.

Hand-coded hairballs of data pipelines

Ad hoc pipelines created without proper practices can result in “hand-coded hairballs,” where complex, intertwined SQL code becomes difficult to manage. For example, a data scientist might create a series of interconnected dbt DAG files with SQL to generate a specific report. Over time, the pipeline becomes a tangled mess as more queries are added without proper documentation or standardization.

This situation poses risks for future maintenance and scalability. Implementing best practices for all types of development, such as modular coding and thorough documentation, can prevent such issues—and reduce your “Tech Debt” as well!

Lakehouse pipeline issues

Poor understanding of ingested data can lead to quality issues in data lakes and data warehouses. For example, raw data dumped into a data lake without proper tagging or metadata can become challenging to use and trust. If the data engineers don’t trust the data or don’t know the data is there, it will be extracted and loaded over and over – causing unnecessary dataset duplication. Therefore users might struggle to find relevant datasets or question the accuracy and completeness of the data.

Implementing robust data governance policies with a full data catalogue would assist in these issues. Tactics such as including metadata management and data quality checks can ensure that data repositories remain useful and reliable sources of information.

Lack of data accountability

Uncoverable issues accumulate when no clear ownership exists, leading to more debt. For example, in a large enterprise with multiple data sources, not having a designated data steward for each dataset can result in neglected data quality issues and outdated information. By no means does a full-fledged data mesh need to be implemented but some love of data product ownership needs to exist.

Assigning clear ownership and accountability for each dataset ensures dedicated personnel are responsible for maintaining data quality and addressing issues promptly.

No Documentation

Poor or missing documentation makes it hard to navigate and maintain data systems. A data warehouse might contain numerous tables with cryptic names and no descriptions (SAP replicated tables for example), making it difficult for new team members to understand the data structure and relationships.

It seems like a small ask but it takes a lot of discipline from the entire data community creating data assets and products. Investing in comprehensive documentation, including data dictionaries and lineage tracking, can significantly improve data accessibility and maintainability.

Addressing data debt requires a structured approach to identify severity, evaluate impact, assess costs, and prioritize remediation efforts. By understanding the implications of data debt and implementing best practices in data governance and collaboration, organizations can mitigate risks, enhance data quality, and improve overall efficiency and productivity.

Regular audits, clear documentation, and fostering a culture of accountability and collaboration are essential steps in managing and reducing data debt. Through these efforts, data professionals can make better decisions because their data assets are valuable and reliable.



06

Business consequences of data debt

The accumulation of data debt brings several significant risks to any organization. These consequences can hinder progress, increase costs, and pose serious compliance issues such as:

01

Decreased productivity and increased costs

Data debt significantly hampers productivity by increasing the time and effort required to handle low-quality data. Data engineers, data scientists, and business analysts often spend excessive hours cleaning, validating, and transforming data before it can be used effectively. This extra work translates into higher operational costs, as more resources are needed to maintain and process the data that data repositories remain useful and reliable sources of information.

02

Poor business decisions and missed opportunities

Poor-quality data leads to unreliable insights, which can result in misguided business decisions. When data is outdated or inaccurate, the analytics and AI models built on it become unreliable. Businesses will start to lose out on opportunities for innovation and growth (often referred to as the “missed opportunity cost”). For instance, marketing campaigns based on incorrect customer data may fail to target the right audience, leading to wasted resources and lost revenue.

03

Regulatory and compliance issues

Data debt can result in severe regulatory compliance issues. For example, if sensitive data like Personally Identifiable Information (PII) or data governed by the General Data Protection Regulation (GDPR) is stored in legacy systems without proper controls, it may be accessed inappropriately. This can lead to breaches of privacy regulations, resulting in substantial fines and damage to the organization’s reputation.

04

Risk of breaking critical pipelines

Legacy data systems often support critical business operations. When changes are made to these systems, there is a risk of breaking data pipelines that other processes rely on. This can cause widespread disruption and halt business activities – requiring significant effort to resolve.

Business consequences of data debt

The pervasive nature of data debt has a pronounced impact on productivity and operational costs. Let's examine this in more detail:

01

Increased time and effort

Data professionals spend excessive time dealing with low-quality data. Identifying, cleaning, and verifying data becomes a resource-intensive process, delaying project timelines and increasing operational costs. A data scientist may spend up to 80% of their time just preparing data for analysis, which leaves little time for actual data modelling and analysis.

02

Hidden Issues

The hidden nature of data debt means that problems often surface unpredictably, leading to unplanned outages and emergency fixes. For example, an unrecorded change in data structure or an undocumented dataset can cause analytics models to fail, requiring urgent troubleshooting and repairs. This unpredictability further strains resources and disrupts planned activities.

03

Legacy system costs

Maintaining outdated data storage systems and appliances, such as on-premises data warehouse appliances or on-premises hadoop clusters, adds to both data debt and technical debt. These systems require specialized knowledge and continuous maintenance, driving up costs and reducing agility. Moreover, as technology evolves, the cost of keeping these systems operational increases, further straining budgets and administrative resources that are becoming harder to find.

04

Poor decision support

Low-quality data hampers effective decision-making. The effort needed to sift through unreliable data delays actionable insights, reducing the responsiveness of the business to market changes and opportunities. For instance, if sales data is not regularly updated and cleaned, forecasts and strategic decisions based on this data will be flawed, potentially leading to missed sales targets.



07

Managing data debt

Prioritize data quality and data management practices

Implement robust data governance policies that emphasize data quality. This includes establishing standards for data entry, regular audits, and continuously monitoring data quality. Educate your organization on data literacy and best practices for data management. Good data management practices should be ingrained in your organizational culture to ensure everyone understands the importance of maintaining high-quality data.

Catalogue all data

Maintain a comprehensive data catalogue, including its location, sensitivity, and ownership. This is critical for large organizations with vast and complex data landscapes. While not a trivial task, it is worth the effort when done properly. A well-maintained data catalogue identifies data assets, and manages data effectively. Use data cataloguing tools to automate the discovery and documentation of data assets, ensuring they are kept up-to-date.

Relate data debt to business impact and ROI

Assess how data debt affects business operations and return on investment. Understanding the business impact of data debt helps prioritize which data assets need immediate attention and resources. For instance, data used in customer-facing applications or critical business processes should be prioritized for cleaning and documentation to avoid potential disruptions.

Ownership and data lineage

Establish clear ownership of data products and maintain detailed data lineage records. Knowing the origin and transformation path of your data allows you to understand the full impact of any changes and ensure responsible data management. Data lineage tools can help track the flow of data from source to destination, providing visibility into how data is transformed and used across the organization.

Collaborate with upstream and downstream teams

Foster communication and collaboration with teams that manage source systems and those that use data products. This ensures alignment and minimizes the risk of disruptions when making changes. Regular meetings and updates between teams can help identify potential issues early and ensure that everyone is aware of changes that may impact their work.

Report on progress and drive accountability

Regularly report on efforts to reduce data debt and hold data owners accountable. Transparency in progress and accountability ensures continuous attention to managing and mitigating data debt. Establish metrics and key performance indicators (KPIs) to track progress, such as the number of datasets documented, the reduction in data errors, and improvements in data quality.

Conclusion

Data debt is an ever-present challenge for data professionals, impacting productivity, costs, and the overall effectiveness of data operations.

By understanding the consequences and employing strategic approaches to manage data debt, organizations can mitigate its adverse effects and harness the full potential of their data assets. Prioritizing data quality, fostering collaboration, and maintaining transparency are key to overcoming data debt and driving business success.

Managing data debt is not a one-time effort but an ongoing process that requires commitment from the entire organization. By investing in good data governance practices, leveraging modern data management tools, and fostering a culture of data quality, enterprises can reduce the burden of data debt and ensure their data assets are reliable, accurate, and valuable.

Interested in learning how to manage your data debt with tools like Matillion? Book a demo to see how Matillion can help reduce data debt and allow teams to be more productive with data.

Or if you'd like to know how big an issue data debt is at your organization, take our free data debt assessment.

Get a demo

Take our data debt assessment