Co-sponsored by:
Hitachi Vantara

BEST PRACTICES REPORT

Q3 2023

Achieving Scalable, Agile, and Comprehensive Data Management and Data Governance

By David Stodder

tdwi TRANSFORMING DATA WITH INTELLIGENCE™

BEST PRACTICES REPORT

Q3 2023

Achieving Scalable, Agile, and Comprehensive Data Management and Data Governance

By David Stodder

Table of Contents

Product and company names mentioned herein may be trademarks and/or registered trademarks of their respective companies. Inclusion of a vendor, product, or service in TDWI research does not constitute an endorsement by TDWI or its management. Sponsorship of a publication should not be construed as an endorsement of the sponsor organization or validation of its claims.

This report is based on independent research and represents TDWI's findings; reader experience may differ. The information contained in this report was obtained from sources believed to be reliable at the time of publication. Features and specifications can and do change frequently; readers are encouraged to visit vendor websites for updated information. TDWI shall not be liable for any omissions or errors in the information in this report.

About the Author

DAVID STODDER is senior director of TDWI Research for business intelligence. He focuses on providing research-based insights and best practices for organizations implementing BI, analytics, data discovery, data visualization, performance management, and related technologies and methods and has been a thought leader in the field for over two decades. Previously, he headed up his own independent firm and served as vice president and research director with Ventana Research. He was the founding chief editor of *Intelligent Enterprise* where he also served as editorial director for nine years. You can reach him by email (dstodder@tdwi.org), on Twitter, and on LinkedIn.

About TDWI Research

TDWI, a division of 1105 Media, Inc., is the premier provider of in-depth, high-quality education and research in the business intelligence and data warehousing industry. TDWI is dedicated to educating business and information technology professionals about the best practices, strategies, techniques, and tools required to successfully design, build, maintain, and enhance business intelligence and data warehousing solutions. TDWI also fosters the advancement of business intelligence and data warehousing research and contributes to knowledge transfer and the professional development of its members. TDWI offers a worldwide membership program, educational conferences, topical educational seminars, role-based training, onsite courses, certification, solution provider partnerships, an awards program for best practices, live webinars, resource-filled publications, an in-depth research program, and a comprehensive website: tdwi.org.

About the TDWI Best Practices Reports Series

This series is designed to educate technical and business professionals about new data and analytics technologies, concepts, or approaches that address a significant problem or issue. Research for the reports is conducted via interviews with industry experts and leading-edge user companies, and it is supplemented by surveys of data professionals.

To support the program, TDWI seeks vendors that collectively wish to evangelize a new approach to solving data and analytics problems or an emerging technology discipline. By banding together, sponsors can validate a new market niche and educate organizations about alternative solutions to critical business intelligence issues. To suggest a topic that meets these requirements, please contact TDWI Senior Research Directors David Stodder (dstodder@tdwi.org), James Kobielus (jkobielus@tdwi.org), or Fern Halper (fhalper@tdwi.org).

Acknowledgments

TDWI would like to thank many people who contributed to this report. First, we appreciate the many users who responded to our survey, especially those who agreed to our requests for phone interviews. Second, our report sponsors, who diligently reviewed outlines, survey questions, and report drafts. Finally, we would like to recognize TDWI's production team: James Powell, Lindsay Stares, Pete Considine, Rod Gosser, and John Bardell.

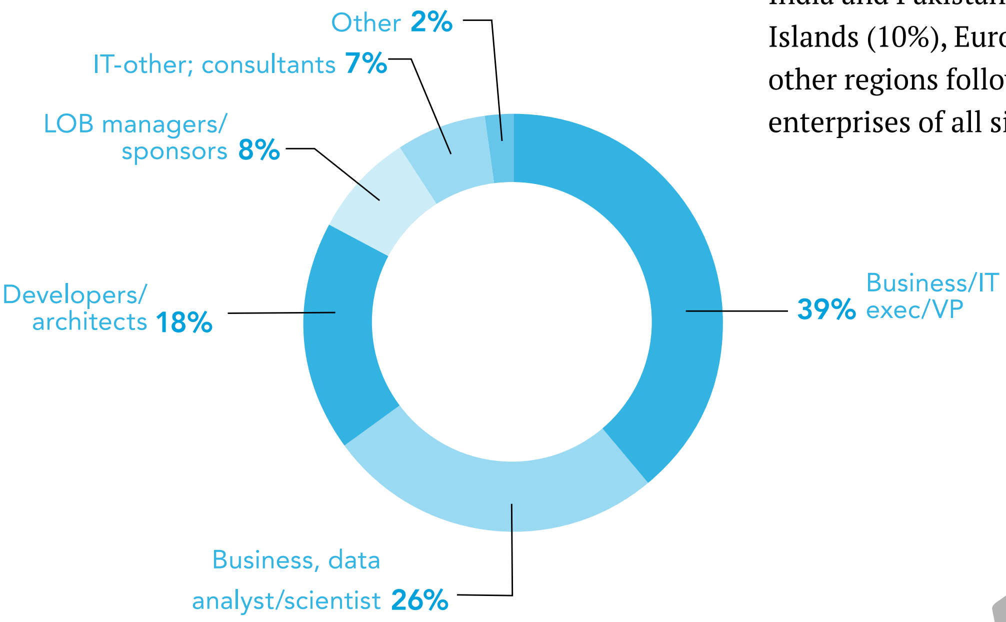
Sponsors

Alation, Denodo, Dataiku, erwin by Quest, Hitachi Vantara, SAP, and Snowflake sponsored the research and writing of this report.

Report purpose. Data demands are great as organizations deploy analytics, artificial intelligence and machine learning (AI/ML), and data-rich applications. Data democratization is multiplying use cases. Organizations face data governance challenges to protect sensitive data and improve trust in data quality. Data management and integration require scalability, speed, and agility. This TDWI Best Practices Report examines priorities and discusses how to overcome challenges to achieve value.

Survey methodology. In June and July 2023, TDWI sent invitations via email to business and IT professionals in our database, asking them to participate in an internet-based survey. The invitation was also posted online and in publications from TDWI and other firms. The survey collected responses from 209 respondents, with 158 completing every question. For this research, all responses are valuable and are included in this report's sample. This explains why the number of respondents varies per question.

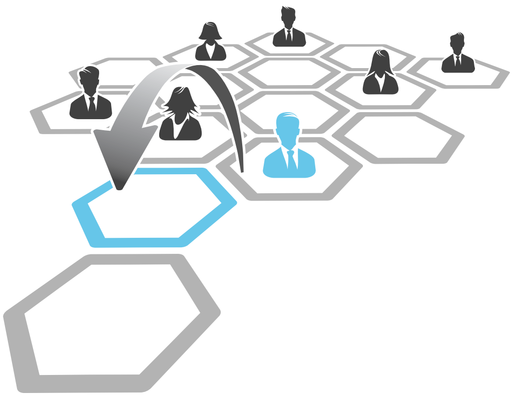
Position



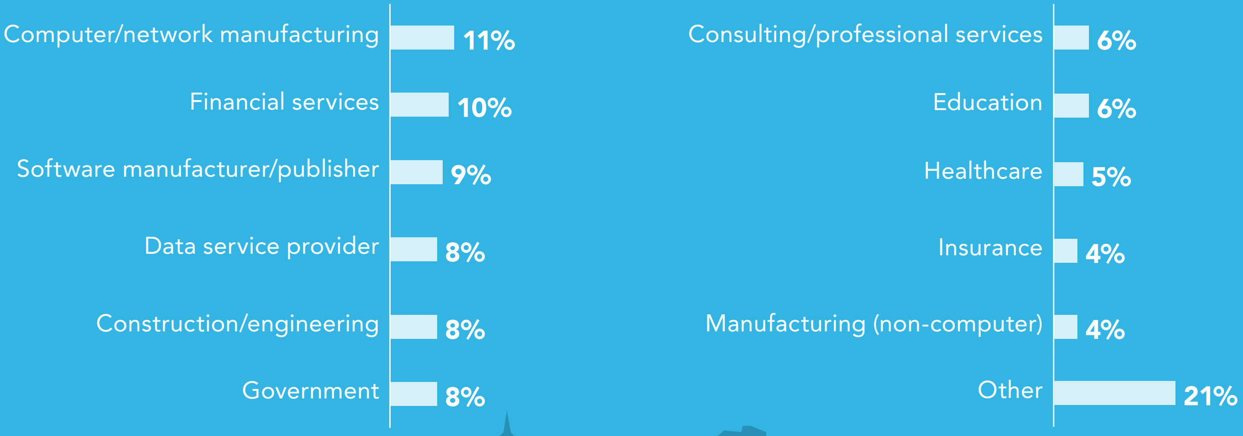
Research methods. In addition to the survey, TDWI conducted interviews with business and IT executives and managers, application developers, and data management and analytics experts. TDWI also received briefings from vendors that offer products and services related to the topics addressed in this report.

Survey demographics. Nearly two in five respondents are business or IT executives and VPs (39%). The second-largest group consists of business or data analysts and data scientists (26%). Third largest is developers and data, application, or enterprise architects (18%). Line-of-business (LOB) managers and business sponsors account for 8% of the respondent population. Other IT staff, consultants, and other titles account for 9% of the total.

Industries vary considerably. Computer/network manufacturing is the largest (11%), followed by financial services (10%), software manufacturing and publishing (9%), data services (8%), construction/engineering and government (8% each), consulting/professional services and education (6% each), healthcare (5%), and insurance (4%). Just under half of respondents reside in the U.S. (48%), with South Asia (primarily India and Pakistan) (25%), Asia and Pacific Islands (10%), Europe (6%), Canada (4%), and other regions following. Respondents come from enterprises of all sizes.

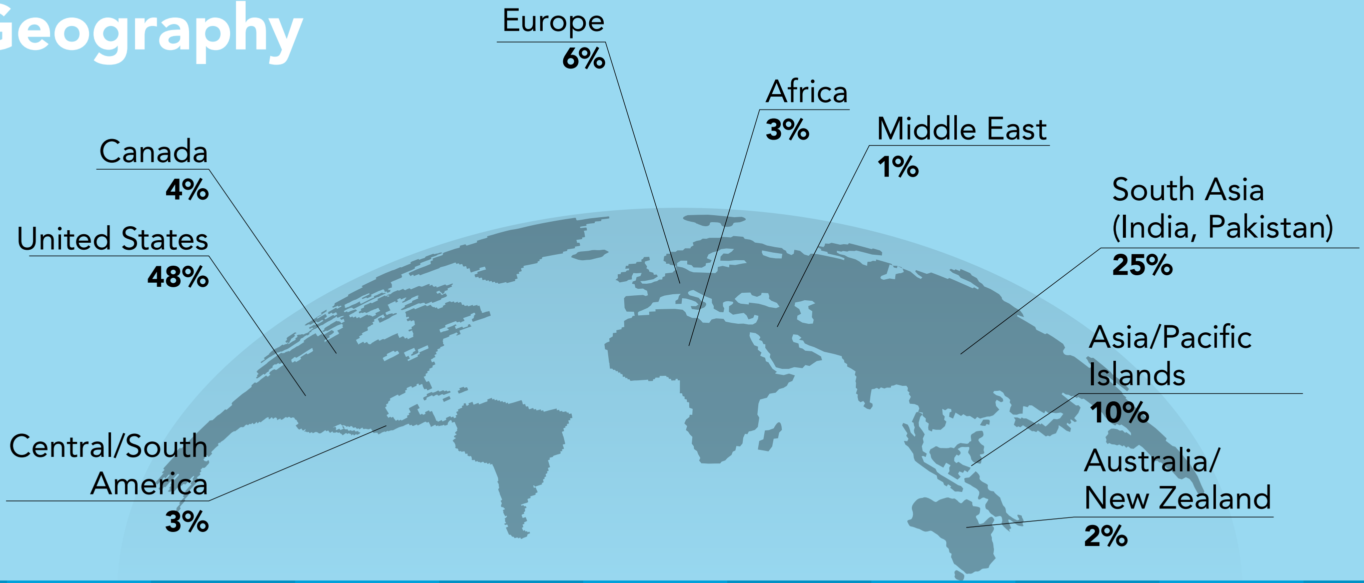


Industry

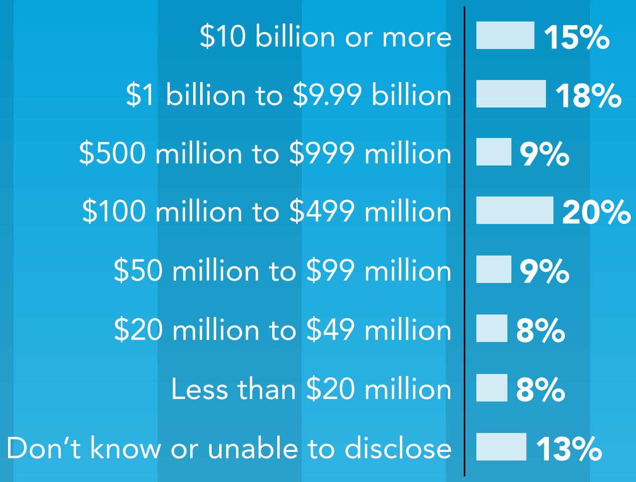


(“Other” consists of multiple industries, each represented by less than 4% of respondents.)

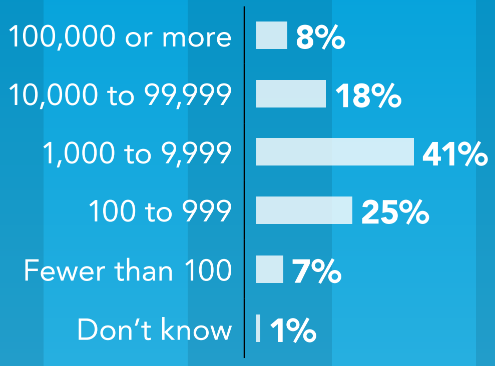
Geography



Company Size by Revenue



Number of Employees



Executive Summary

The data explosion continues to accelerate across distributed landscapes with data on premises and on multiple cloud data platforms. Organizations face challenges as well as tremendous potential for increasing the value of data assets, including through data monetization—potential that can go untapped without good data management and governance. Most organizations have a democratized spectrum of users, creating ever-greater demand for data inside and outside their organizations. This drives the need for diverse types of data and different levels of data timeliness, as well as differing requirements for data governance, integration, and quality.

This TDWI Best Practices Report focuses on understanding current challenges and providing best practices insights for modernizing processes and deploying technologies to solve them. Analytics workloads augmented with AI/ML are critical to competing in every industry. Data-driven business initiatives depend on scalable, agile, and comprehensive data management and governance. The latest applications embed sophisticated analytics using AI/ML capabilities that must be provisioned with continuous, integrated, curated data to deliver insights to all users. Flexibility is key to keeping pace with business demand and unanticipated events.

Organizations are advancing with AI/ML through easier, automated capabilities such as autoML. Some are investigating large language models (LLMs) and generative AI. To move forward, organizations need to modernize data management, integration, and governance and align investments with evolving business requirements. Legacy technologies and practices often force data scientists, data and business analysts, and business users to spend too much time acquiring, integrating, and preparing data. We discuss how AI-infused automation in data integration and preparation processes are maturing to enable users to focus more time on solving business challenges and achieving data-driven innovation.

> *Our research finds that most organizations have only isolated success in managing and governing data to meet objectives.*

Our research finds that most organizations have only isolated success in managing and governing data to meet objectives. Accelerating growth in data volume, workloads, and users across distributed and disparate data landscapes generates pressure that can lead to chaos and higher costs. Our report discusses how organizations can improve the balance between enterprise data governance and the agility required for self-service user empowerment.

Limited data access is a problem when organizations are trying to develop new insights about concerns such as customer behavior, supply chains, public health, operational cost drivers, and business performance. Distributed data dispersed across silos is a major challenge to gaining complete views of all data about these concerns. It also presents challenges to holistic data governance and management. Research in this report shows many enterprises now have experience with or plans for bridging distributed data through data virtualization, data mesh, and data fabric architectures or consolidating disparate data into a unified cloud data platform.

All data architectures today rely on technology modernization to capture and manage metadata and other knowledge about all the data, including data lineage. Organizations are expanding use of data catalogs and additional data intelligence and semantic layer systems. This report discusses current satisfaction with data catalogs, business glossaries, and metadata management systems and where organizations need to improve to increase satisfaction.

The report concludes with a discussion of how modern technologies and practices are coming together to create unified data environments. It discusses the importance of making this unity flexible rather than restrictive. Finding the right balance enables organizations to empower teams to maximize the value of enterprise data assets. We close the report with 10 recommended best practices for success.

Modernizing the Data Foundation for Business Insights

Opportunities to drive higher value from data and analytics can be squandered if organizations lack scalable, agile, and comprehensive data management and governance. Data democratization is accelerating demands in organizations of all sizes and across all industries. Users require trusted data for strategic forecasting, risk assessment, and customer engagement as well as to drive daily decisions. Data fuels simple operational dashboards as well as complex analytics projects augmented by AI/ML. Automated decisions in data-driven applications embedded with analytics and AI/ML require continuous access to trusted data for real-time response to customer behavior or business events and for pattern detection.

> *Organizations cannot advance if they stand pat with inadequate legacy tools and platforms when data environments are constantly changing.*

Data management, data integration, and data governance technologies and practices together form the foundation for achieving ambitious objectives with data. Organizations cannot advance if they stand pat with inadequate legacy tools and platforms when data environments are constantly changing. Agility is essential for responding to often unanticipated changes in business requirements and conditions. Organizations need to continuously evolve data management and data governance to overcome obstacles that prevent the right people and applications from accessing the right data at the right time.

This TDWI Best Practices Report focuses on trends, challenges, and opportunities to modernize data management, data integration, and data governance technologies and practices. Updating data platforms, distributed data integration layers, and tools for data management and data governance is essential for supporting growth in the use of data and increasing efficiency and satisfaction. This report will examine current experiences and practices and discuss how to address evolving challenges.

How successful are organizations currently in managing and governing data, especially to support analytics and AI/ML? The prevailing answer, according to our research, is “it depends.” Most organizations we surveyed experience inconsistency; 38% say they have isolated areas of success and agility, but capabilities across their organizations are uneven (figure not shown). Only 14% say that they are highly successful and confident in how they manage and govern data today and can respond to change. Twice that percentage (28%) indicate that although they regard their organizations' current data management and data governance as successful, they anticipate challenges in responding to change.

Fortunately, only a small percentage of respondents indicate a total lack of confidence; just 8% agree with the statement that they are "struggling to meet current requirements and are not well prepared for change." However, nearly as many in our research (10%) say that they are just getting started with managing and governing data for more than one project, application, or team.

Inexperience is often the case with small and midsize businesses (SMBs). As they grow, SMBs often bump up against the limitations of inexpensive and readily available tools such as spreadsheets; hand-coded, point-to-point data pipelines; and siloed data platforms. As data grows in volume and variety and more users attempt complex analytics, these users spend too much time on ingesting, extracting, and preparing data. Redundant or inconsistent data pipelines and transformation processes increase costs and lead to unsatisfactory performance. Decision makers are reluctant to trust the resulting data findings. Risks of sensitive data exposure are everywhere, demanding more holistic data governance to at least ensure adherence to data privacy regulations.

Managing and Governing Data for Business Objectives

When traditional IT systems and practices are isolated from business functions, it is not always clear how data management and governance relate to the achievement of business objectives. They can seem like fixed services; organizations that do not consider themselves "in the IT business" are mostly focused on reducing IT costs. However, business-driven adoption of cloud computing and trends toward funding some or all data management through operations rather than as standard capital expenditures have put increased emphasis on aligning resource utilization with business operations and objectives.

> *Usage-based service pricing models offered today by cloud data platform, application, and tool providers make it easier to see the relationship between resource consumption and business activity.*

Usage-based service pricing models offered today by cloud data platform, application, and tool providers make it easier to see the relationship between resource consumption and business activity. With these models, business functions can understand how business activities drive consumption of storage, processing, data management, and data integration services. With holistic data observability monitoring, organizations have visibility into how bottlenecks slowing data access—such as having to wait for high-volume data movement or slow data transformation processes—ultimately affect operational performance, customer satisfaction, and the ability to achieve business objectives.

Cloud service monitoring tools play an essential role in tracking usage so companies can avoid paying for unused or underutilized resources. Organizations can deploy metrics and data observability capabilities, which will be discussed in more detail later, to understand usage patterns and to know, for example, which data sets, reports, dashboards, analytics models, data warehouses, and data applications are delivering the most value.

Monitoring and metrics enable organizations to analyze usage data to determine where modernization, such as greater automation, would be most beneficial. Monitoring visibility also helps to evaluate whether resources are being devoted to unimportant projects or where your organization needs to improve access to underutilized data assets. With analytics augmented by AI/ML (including generative AI and LLMs) rising in importance, organizations must ensure that they have resources and governance practices in place as they scale up in numbers of users, workloads, and data volumes.

Understanding how investments contribute to business objectives is critical. To drill down deeper into the contribution of data management and governance investments to business success, we asked research participants to indicate how successfully their organizations are managing and governing data and realizing value through analytics, including AI/ML, to achieve objectives across business functions and initiatives. In Figure 1, we can see that organizations surveyed are most successful at meeting objectives important to regulatory compliance and protecting data; 30% are very successful and 43% are somewhat successful. These results show the importance organizations place on succeeding with core data governance priorities for regulatory compliance, often led by enterprise IT.

Organizations indicate the most success in three additional areas:

- **Customer experience.** Big data analytics is revolutionizing how organizations personalize marketing offers and shape customer experiences. The volume of behavioral data generated by customer activity across channels can give marketers a tremendous resource for gaining a detailed view of customers' historical and real-time behavior, which organizations can analyze to generate predictive insights into how customers will respond to future offers. Figure 1 shows that 28% are very successful and 43% are somewhat successful with data management and governance for improving customer experiences.

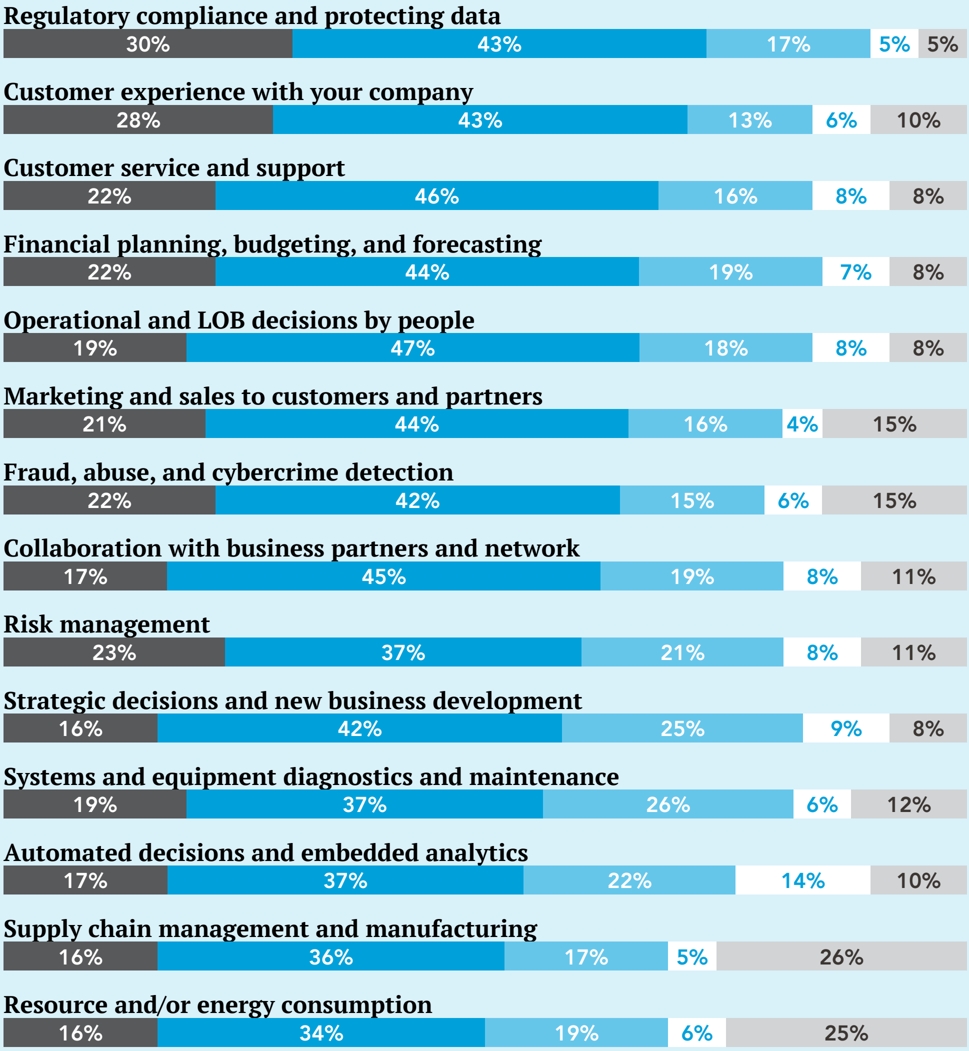
- **Customer service and support.** Integrating data to gain 360-degree views of customers is valuable for all customer-centric objectives, including excelling in service and support. Valuable data often includes semi- and unstructured customer service records, call center interaction records, customer satisfaction survey data, and information generated by online behavior and engagement. Figure 1 shows that 22% are very successful and 46% are somewhat successful with data management and governance for improving customer service and support.

- **Financial planning, budgeting, and forecasting.** Organizations are reasonably satisfied with their data management and governance for these activities; 22% are very successful and 44% somewhat successful. Users depend on data management and governance to provision trusted, high-quality data and enable timely single views drawn from multiple sources for accurate snapshots and forecasting decisions. Agility is an important attribute for organizations to update plans and budgets as conditions change.

Figure 1

To achieve business objectives in the following functions and initiatives, how successful is your organization in managing and governing data and realizing value through analytics, including AI/ML?

Based on answers from 209 respondents. Ordered by combined "very successful" and "somewhat successful" responses.



When we combine "somewhat unsuccessful" and "not at all successful" responses, Figure 1 shows that organizations are less successful in these three areas:

- **Automated decisions and embedded analytics.** Cutting-edge applications embed analytics to detect events, anomalies, and patterns and provide users with recommendations for action. Some use analytics to drive automated decisions. The research indicates that many organizations are not yet fully satisfied with data management and governance for supporting these capabilities, with 22% somewhat unsuccessful and 14% not at all successful.

- **Strategic decisions and new business development.** When organizations are undergoing significant, sometimes unexpected changes in markets, business networks, and economic environments, they need agile data management and governance. The research indicates some frustration with support for strategic decision-making and new business development, with 25% somewhat unsuccessful and 9% not at all successful.

- **Systems and equipment diagnostics and maintenance.** Analytics augmented with AI/ML is revolutionizing these functions, enabling organizations to develop predictive and prescriptive algorithms to know when machinery is likely to break down. They can prepare maintenance and repair based on data rather than traditional schedules. Data management, integration, and governance are critical to gathering and preparing data streamed from end points and enabling integrated views of historical and real-time data. About one quarter (26%) are somewhat unsuccessful in supporting these advanced capabilities and 6% are not at all successful.

Modernization Drivers

Organizations are updating existing data platforms or migrating to new ones and using modern data integration, data management, and governance practices and systems to achieve greater success with the objectives shown in Figure 1. To gain insight into current modernization priorities, we asked research participants to identify their organizations' most significant drivers (see Figure 2).

> *The largest percentage of respondents (50%) say that improving trust in data quality, accuracy, and completeness is one of their most significant drivers.*

The top two choices demonstrate the importance of addressing data governance as part of modernization. We will discuss specific data governance issues in depth in the next section, but here we gain a broader view of the relative importance of various drivers and priorities. Half of those surveyed (50%) say that improving trust in data quality, accuracy, and completeness is one of their most significant drivers. This issue is the focus of what many in the industry call "offensive" data governance. Protecting sensitive data and preventing unauthorized use and sharing, regarded as "defensive" data governance, ranks second, with 38% identifying it as one of their top three drivers. Just under one-third say that complying with regulations and enabling regulatory reporting is a significant driver.

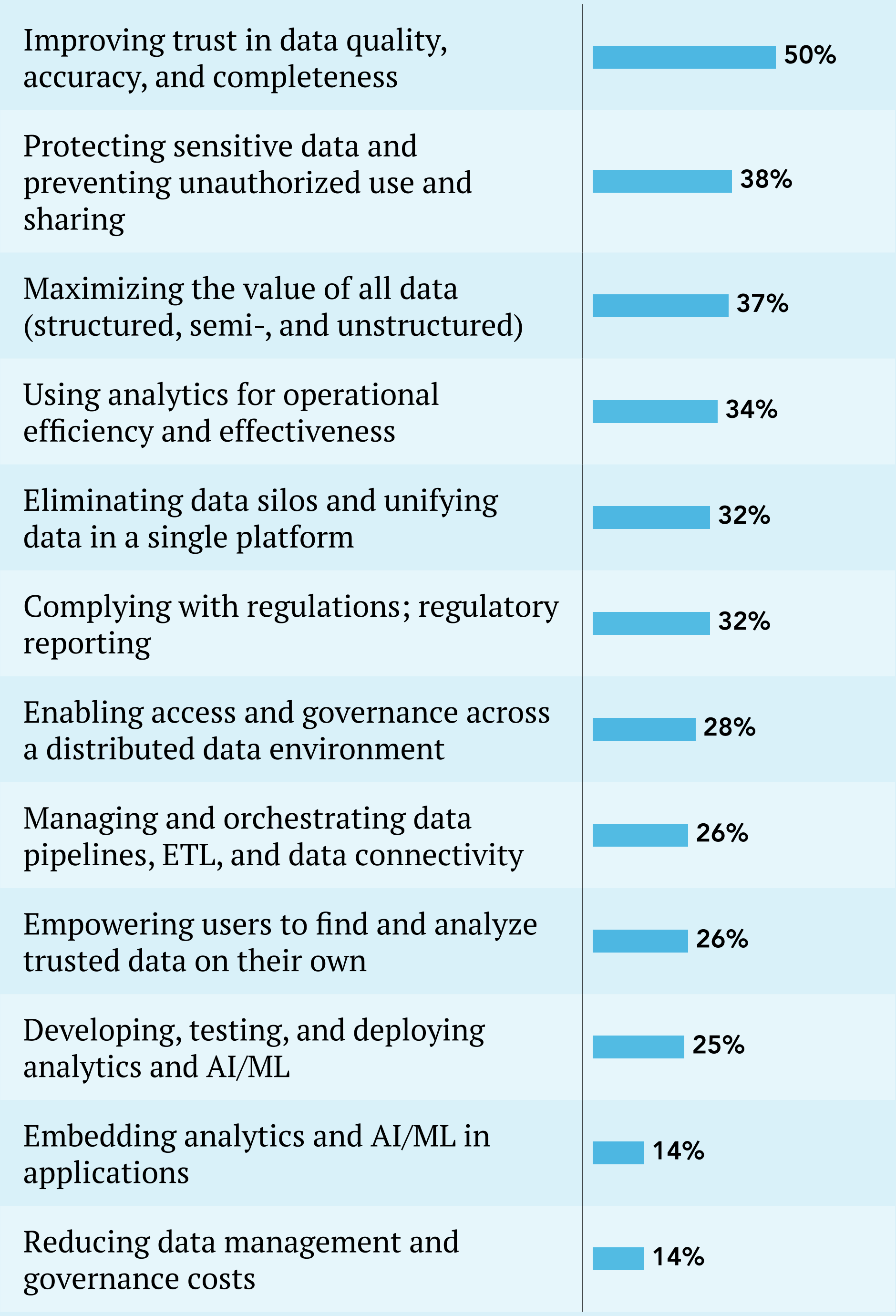
Both offensive and defensive data governance are essential to maximizing the value of all data, whether structured, semistructured, or unstructured. Figure 2 shows that this ranks third among drivers with 37% choosing it as one of their top three. This selection also indicates the importance of collecting and curating diverse data for analytics and AI/ML development. Data lakes, unified data lakehouse platforms that integrate data warehouses and data lakes, and NoSQL systems have become critical technologies for managing diverse data. Many organizations are focusing on addressing distributed data challenges; eliminating data silos and unifying them in a single platform ranks fifth, with 32% saying it is one of their top drivers.

Analytics and AI/ML development and operationalization are significant drivers. Ranked fourth by participants is the usage of analytics for operational efficiency and effectiveness. Organizations want to enable managers in business functions and lines of business (LOBs) to examine different perspectives on data used in performance metrics and dashboards to determine, for example, the best ways to reduce unnecessary costs and delays.

Figure 2

What are the most significant drivers behind your organization's current efforts and plans to modernize data management and governance?

Based on answers from 196 respondents, who were asked to select at least their top three objectives.



Nearly half of research participants (48%, not shown) who indicated earlier that their organizations are highly successful and confident in how they manage and govern data say that using analytics for operational efficiency and effectiveness is one of their most significant drivers. Among this selection of research participants, 72% say that developing, testing, and deploying analytics and AI/ML is a top driver (compared to 25% of all respondents saying this driver is significant).

Satisfying Analytics and AI/ML Requirements

The importance of analytics and AI/ML demands that organizations develop a comprehensive strategy for addressing projects' data management, integration, and governance requirements. Uncoordinated, piecemeal approaches tend to increase challenges as well as costs. However, rather than constrain analytics, data science, and data engineering teams with too much centralized control, many organizations are seeking to empower teams so they have the ability and independence to develop sharable data products and analytics that meet their domain's business objectives. Later, this report will discuss strategies such as the data mesh for finding the right balance between centralized governance and self-service empowerment.

> *Many organizations are seeking to empower teams so they have the ability and independence to develop sharable data products and analytics that meet their domain's business objectives.*

Because they are part of the foundation, addressing pain points in data management and data governance is vital to moving forward with analytics and AI/ML. We asked research participants for their data management and governance improvement priorities specifically for analytics and AI/ML. In Figure 3, we can see that reducing time and costs associated with data collection and preparation is the highest priority for the largest percentage of respondents (43%). Increasing data availability for AI model development, training, and testing is a close second (40%). Additionally, nearly one-quarter say that provisioning data for feature engineering is a major focus (22%).

Nearly one-third say that making it easier for users to operationalize data pipelines and transformation is a top priority (32%). A significant percentage prioritize increasing automation in data pipelines and AI/ML workflows (29%). These and additional results in Figure 3 suggest that data problems are hindering analytics and AI/ML projects and solving them could accelerate value.

> *Organizations need visibility to detect data changes after an AI model has been deployed so they can understand the reasons for any poor model performance.*

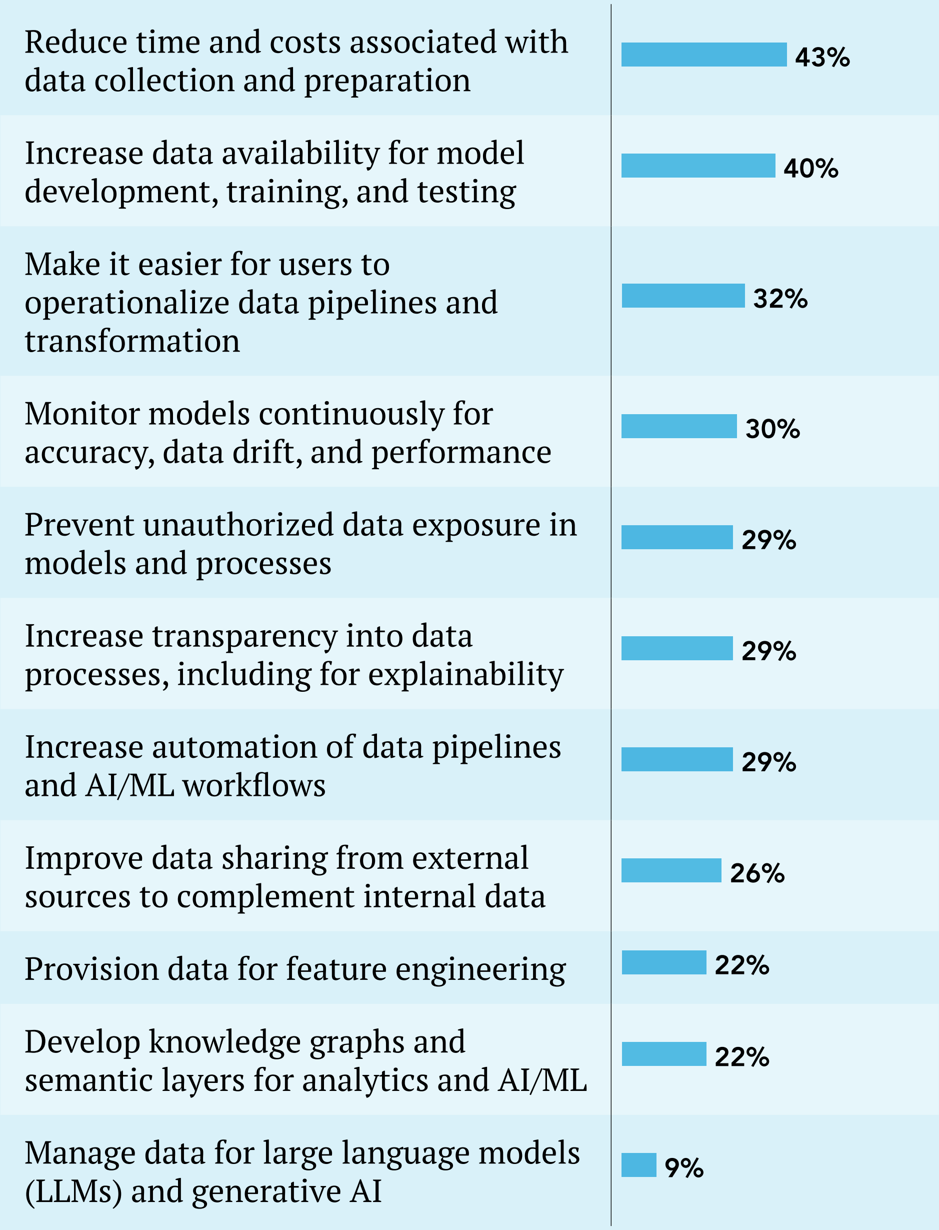
Strong interest exists in extending data governance to AI model governance. In Figure 3, nearly one-third say that monitoring models continuously for accuracy, data drift, and performance is a top priority (30%). Organizations need such visibility to detect data changes after an AI model has been deployed so they can understand the reasons for any poor model performance. Nearly as many prioritize defensive model governance through prevention of unauthorized data exposure in models and processes (29%).

Explainability is often considered a key element of AI model governance; 29% say they prioritize better transparency into data processing, including for explainability. Regulatory compliance is a major driver behind explainability, which is the ability to explain business decisions that are based on a model, such as denial of credit to a customer. In addition to regulatory compliance, however, organizations also need explainability to improve decision-makers' trust in AI models and to know how to tune model performance for better results.

Figure 3

To improve data management and governance for analytics and AI/ML specifically, which of the following objectives are your organization's highest priority?

Based on answers from 184 respondents, who were asked to select at least their top three objectives.



Data Governance for Protecting Data and Driving Value

Modern data governance fuses two critical sets of objectives. As noted earlier, the defensive objectives focus on articulating and implementing rules and policies for protecting sensitive data such as customers' personally identifiable information (PII) from unauthorized access and sharing. Organizations need to comply with legal regulations as well as industry and internal policies; they need the ability to respond to audits to demonstrate compliance. Users must know and adhere to rules and policies as they access and interact with data.

Offensive data governance objectives aim at creating business value through increased organizational understanding and trust in the data and its fitness for each purpose. Users need to be able to easily discover available data and be confident in the data they choose to consume and share in reports, metrics, dashboards, and notifications. If people lose trust in the data—for example, by depending on a chart that turns out to contain inaccurate or incomplete data—they will be reluctant to turn to data again. Analytics models built and tested with poor and incomplete data can lead to bad business decisions and unsatisfactory customer interactions.

Thus, improving the value of data assets through a data curation process focused on discoverability, data quality, accuracy, consistency, and completeness is vital. Some organizations integrate offensive data governance with programs for improving user data literacy. Data literacy addresses primarily human aspects of how people interact with data. The primary goal is to raise individuals' proficiency with understanding what data means and their ability to communicate and share analytics insights.

Overall, most research participants have a positive view of how their organizations are achieving defensive and offensive data governance objectives. More than half (54%) regard their organizations as somewhat successful and 19% say they are very successful (see Figure 4). The percentages are consistent across all sizes of organizations surveyed. Demonstrating the importance of data governance to growth in analytics and AI/ML, 43% of respondents who say their organization is very successful indicate that developing, testing, and deploying analytics and AI/ML is their top data management and governance modernization driver (not shown).

Meeting Compliance and Data Protection Priorities

As organizations grow more reliant on data for decisions and applications processes, data volumes and the number and complexity of data-intensive workloads make both defensive and offensive data governance more challenging. Looking first at defensive data governance, Figure 5 offers a fuller view of where organizations are most and least successful in addressing regulatory compliance and sensitive data protection priorities.

> *Almost two-thirds indicate success (64%) in training to ensure users understand guardrails, establishing data stewardship roles, and using modern tools that can automate and embed data governance constraints.*

Here are areas where organizations are most successful:

- **Complying with data privacy and other industry regulations.** The imperative of acting in compliance with regulations and avoiding penalties understandably gets the most attention. Nearly one-third say they are very successful (30%) and more than half are somewhat successful (55%).

- **Controlling access to sensitive data such as PII.** Access controls supported by tools and up-to-date policies are critical to protecting sensitive data. Distributed data landscapes such as hybrid multicloud environments are a challenge for controlling access; some organizations use a data fabric or data virtualization layer to create a central point of control. Organizations show measured success in controlling access, with 30% very successful and 50% somewhat successful.

- **Setting data governance rules and policies and keeping them up to date.** Data environments are constantly changing as organizations collect new data and democratize access. This makes it critical to review rules and policies and ensure they remain accurate and relevant. In our research, nearly three-quarters of organizations (73%) are successful to some degree, with 31% very successful.

- **Training users in governance and responsible data use.** Training users so they know their responsibilities and accountability for compliance is important. In Figure 5, we can see that most organizations surveyed regard themselves as successful in training users (64%), although just 21% say they are very successful and 30% are either somewhat unsuccessful or not at all successful, indicating room for improvement.

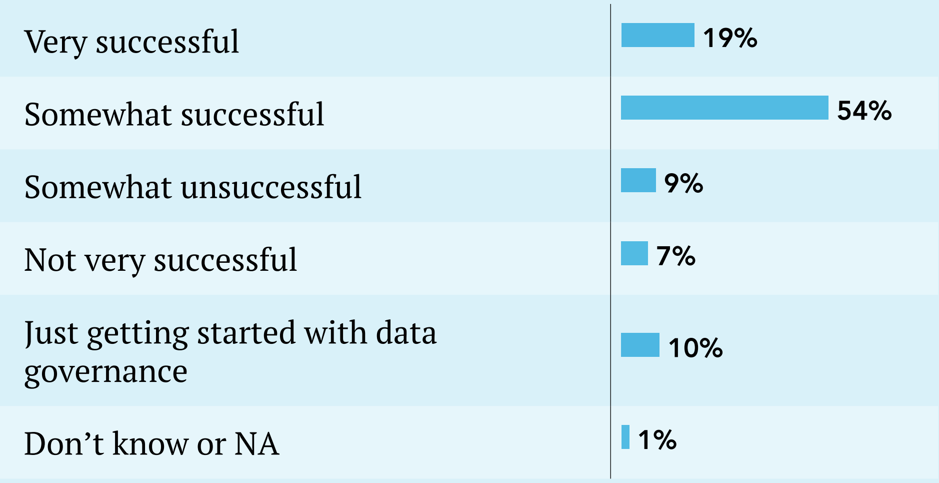
- **Ensuring data governance in self-service and analytics.** As organizations democratize data access and analytics, balancing users' needs with data governance priorities becomes challenging. Training to ensure users understand guardrails, establishing data stewardship roles for guidance and mentoring, and using modern tools that can automate and embed data governance constraints are all critical to success. Almost two-thirds of respondents indicate success (64%). However, only 20% say their organizations are very successful and 30% say they are either somewhat unsuccessful or not at all successful. About one-fifth (21%) say finding the right balance between self-service data access and enterprise data governance is one of their top challenges (not shown).

The research shows that organizations are least successful in improving discoverability and having an accurate inventory of sensitive data, which is a requirement for most data privacy regulations including GDPR. One-quarter of respondents say their organizations are somewhat unsuccessful

Figure 4

Overall, how successful is your organization currently with its data governance for ensuring regulatory compliance, protecting sensitive data, and improving trust in the data people use and share?

Based on answers from 183 respondents.



(25%) and 10% are not at all successful. We will discuss later how data catalogs and similar metadata management systems are valuable for knowing the location of sensitive data and establishing an inventory.

> *The research shows that organizations are least successful in improving discoverability and having an accurate inventory of sensitive data.*

Other issues that organizations indicate are challenging include:

- **Coordinating data governance with data security and use of masking techniques.** One-third of organizations surveyed say they are unsuccessful in establishing good coordination (33%), with 10% reporting that they are not at all successful. Governance committees should bring leaders together to ensure alignment between data governance and data security measures such as masking, anonymizing and de-identifying data. Regulations often have conditions regarding acceptable use of these techniques for certain types of customer data. Organizations should evaluate solutions such as automated tag-based masking capabilities in data platforms and tools that offer flexibility in whether your organization assigns masking to databases, tables, schema, or columns.

- **Embedding governance constraints in pipelines, tools, and applications.** One-third of organizations surveyed are unsuccessful in implementing embedded constraints (33%), with 12% indicating they are not at all successful. This priority involves using embedded capabilities that may not be available in legacy data integration and management systems. Traditional practices often require users to consult policy documentation to determine whether their data use and sharing complies with governance policies, which can lead to inconsistent compliance.

- **Monitoring governance during data loading, movement, and replication.** Organizations need to monitor exposure risks not only while data assets are at rest in data lakes and data warehouses but also while they are in motion during data integration or cloud migration. Governing data in motion to avoid exposure appears to be challenging for some organizations; 22% say they are somewhat unsuccessful and 11% are not at all successful. Leading data integration tools and data platforms offer automated capabilities for monitoring data governance during loading, movement, and replication.

Data Governance for Improving Data Trust

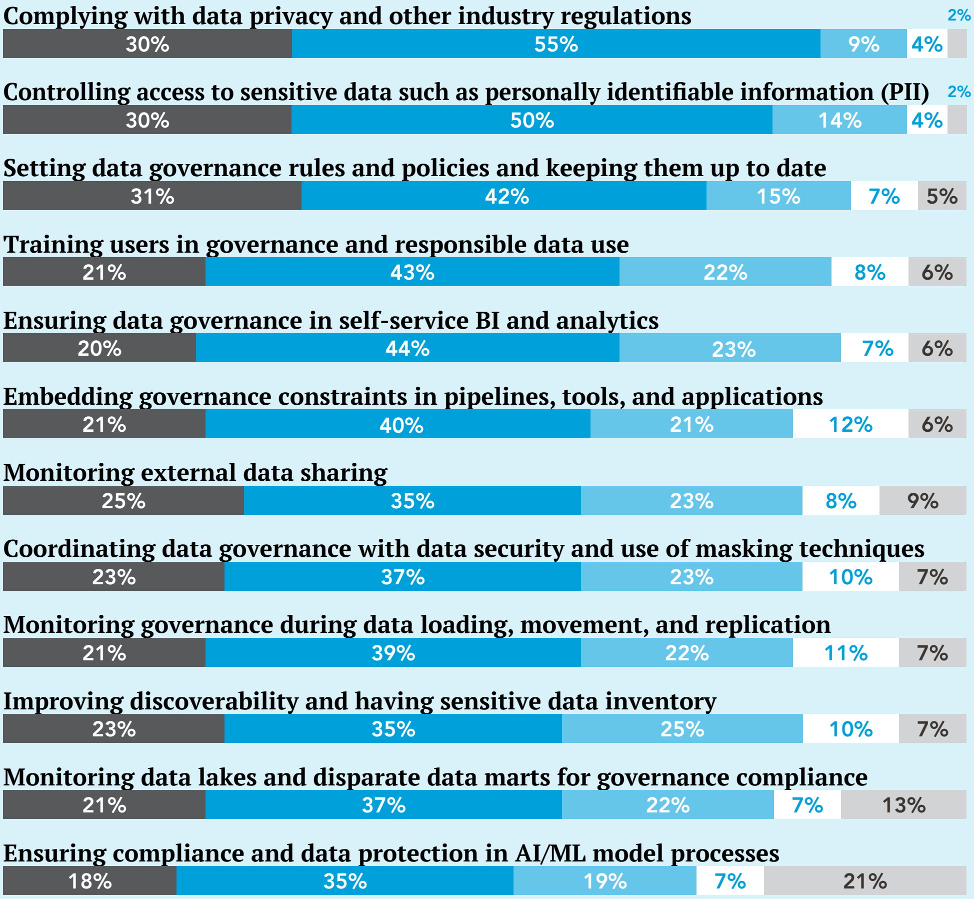
The offensive side of data governance is about building data trust. Data curation steps for improving the data's quality, validity, integrity, and authenticity are central to this objective. Integrating data governance and data trust is how organizations can avoid a "Wild West" in which the expansion in self-service BI and analytics happens without proper data governance guardrails. Through data governance, you can clarify who is responsible for data authoring and data curation.

Data quality monitoring is valuable for ensuring requisite profiling, cleansing, validation, and enrichment, as well as uncovering problems

Figure 5

Regarding regulatory compliance and protecting sensitive data, how successful is your organization with its data governance for addressing the following priorities?

Based on answers from 177 respondents. Ordered by combined "very successful" and "somewhat successful" responses.



such as data inconsistency and redundancy. Data governance policies can indicate when data owners or other experts should be notified about problems with the data. Modern tools can automate monitoring. AI-infused automation is playing an increasing role in data curation processes, enabling organizations to scale up to higher and more diverse data volumes faster and spot anomalies that require attention.

TDWI research asked which of the actions listed in Figure 6 are being undertaken within respondents' organizations to increase users' trust in data quality, including its accuracy and completeness. The most common action is to validate new data and address data quality issues at sources (50%). Organizations can use data intelligence tools and processes such as data catalogs to know more about source data quality. Scalability is often a challenge when sources become numerous and voluminous.

The second most common action is to monitor data quality and ensure fit for each purpose (45%). With the spectrum of use cases in most organizations stretching from simple data consumption to deep analytics and AI/ML, it is important to target data curation carefully. Data quality monitoring and processes need to be fit for each purpose. AI-augmented tools for automatically detecting data quality issues are evolving to make it easier for users to learn about data quality and address problems; 25% say they are currently deploying these tools and 16% are automating data quality recommendations to users.

> *Data stewards can oversee data quality processes, manage metadata and master data, and guide users to data that is fit for purpose. Automated data intelligence tools are essential.*

Data stewardship is a priority. Almost half of organizations (45%) say that enhancing data stewardship to include data quality and trust is an action they are taking. Data stewardship is valuable for both defensive and offensive data governance. Along with being experts in regulatory compliance, data stewards can oversee data quality processes, manage metadata and master data, and guide users to data that is fit for purpose. Automated data intelligence tools such as data catalogs (or similar functionality embedded within some data intelligence tools and data platforms), as well as embedded data governance constraints and data governance workflows, are essential to data stewards because they reduce manual work, offer more complete information about data, and streamline updates to metadata as the data environment changes.

Tracking data lineage is also being undertaken by 45% of organizations surveyed. Data lineage information includes visual documentation of the data sources, ownership, and transformations throughout an organization's data landscape. Some rely on data lineage–specific tools. Others leverage data catalogs that can track data lineage and serve as a central repository for information, combining visibility of data quality scoring and data classification as well. Along with enabling better defensive data governance and managing the inventory of sensitive data, data lineage information makes it easier for self-service business users to find and contextually understand high-value, relevant, and trusted data.

Many organizations employ a data catalog for data governance. Our research shows that 42% use a data catalog to inventory data and identify data quality issues. Automation in modern data catalogs is critical to meeting scalability and data diversity challenges. As noted, data catalogs can help organizations document data inventories as required by regulations and data lineage. Easier-to-use interfaces in modern data catalogs enable different types of users to access data intelligence and determine whether they can trust data to be fit for their needs.

Top Data Governance Challenges

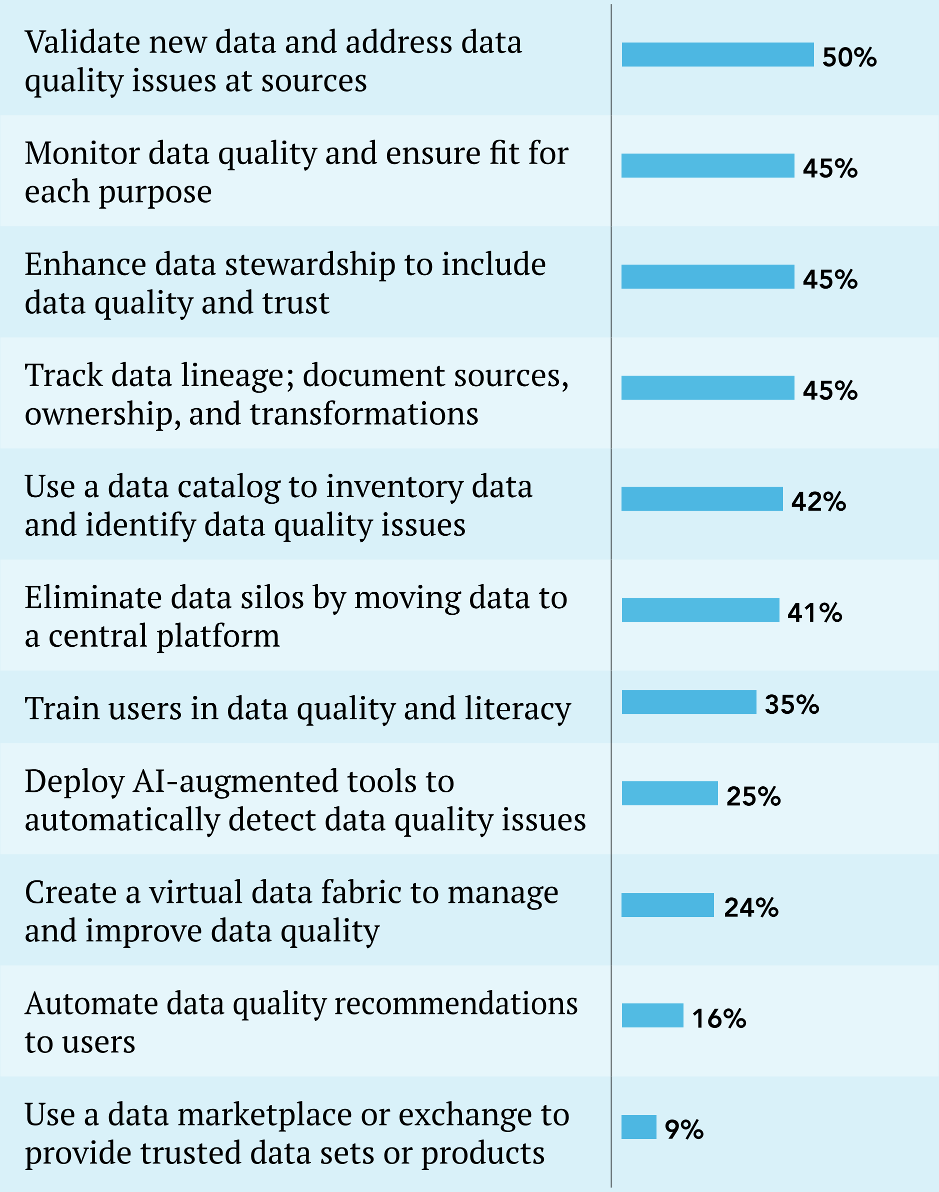
We close this section of the report with a look at the most pressing data governance challenges confronting organizations in our research. These fall into the following areas:

Governing distributed data across silos. Data distributed across a hybrid multicloud environment makes it difficult to improve data quality, track data lineage, and monitor data use for protecting sensitive data. Data silos, including independent data marts and data lakes, fragment data across systems; 41% cite data silos as one of their top three challenges (figure not shown for research reported in this section). Silo creation often accelerates with data democratization as project teams or individual users set up their own data repositories.

Figure 6

To increase users' trust in the data quality, including accuracy and completeness, which of the following actions are being undertaken within your organization?

Based on answers from 177 respondents, who could select all that applied.



> *One-third (33%) say that unifying data governance across a distributed data architecture is a major challenge.*

One-third (33%) say that unifying data governance across a distributed data architecture is a major challenge and 28% say it is difficult to ensure data governance across on-premises and cloud-based data. Governance challenges are motivating a significant percentage of organizations surveyed (41%) to eliminate data silos by moving data to a central data platform. Some are instead choosing to create a virtual data fabric to improve holistic data governance, data quality, and data management across a distributed data environment (24%). These options will be discussed in more detail in sections to follow.

Integrating data governance with data integration and data pipelines. Organizations should ensure that data governance is part of processes for ingesting, collecting, and preparing data, not just once the data is at rest in storage and managed as part of a data warehouse or data lake. Nearly one third (30%) say that ensuring data trust and protection in data pipelines, data loading, and extract, transform, and load (ETL) processes is one of their top challenges. Organizations should employ modern, automated capabilities embedded in data integration tools and platforms. Allowing data pipeline processes to automatically access a data catalog or other metadata management will enable users to find trusted data faster and help data stewards keep track of sensitive data.

Addressing people and process issues. Earlier, our research noted that most organizations are confident in their ability to set governance rules and policies and keep them up to date. Yet, a significant percentage (40%) say that ensuring clear rules, policies, and responsibilities for data use is an ongoing challenge. This is because data landscapes are constantly changing, as are data privacy regulations and the dangers of exposure through errors and abuse, such as hacking. Organizations should continuously monitor the effectiveness of rules and policies so that they are effective and up to date.

As we noted earlier, integration of data governance with analytics model governance is an emerging priority. In our research, 14% of organizations surveyed regard this as one of their top challenges. Organizations should facilitate communication between data governance and data science teams to align policies and processes.

> *Automated tools and data catalogs are critical to reducing hands-on monitoring and manual policy enforcement so data stewards can work efficiently.*

Many organizations surveyed say that finding the right people to be data stewards is a primary challenge (38%). Organizations often struggle to identify data stewards, most of whom have “day jobs” as business subject matter experts (SMEs), data analysts, business analysts, or data warehouse managers. Automated tools and data catalogs are critical to reducing hands-on monitoring and manual policy enforcement so data stewards can work efficiently. By collecting information about data activity among users, data knowledge, data ownership, and relevant data governance constraints, data catalogs can help data stewards avoid having to track down disparate information.

Using automation effectively. More than one quarter of organizations surveyed (28%) say that automating data governance, data quality, and data stewardship tasks is one of their top challenges. The same percentage need solutions that will enable them to increase automated scanning and tagging of new data sets (28%). Organizations should evaluate forward-looking tools and platforms that provide automation of these and similar functions.

Data Management for Expanding Opportunities

Pressures on data management continue to rise as data grows in volume, variety, and velocity. Rather than rely on a traditional monolithic system, organizations increasingly rely on a portfolio of data management solutions to support access to the range of historical, continuously updated, and real-time data for BI, analytics, and AI/ML. To maximize the value of data assets, including through monetization and participation in data marketplaces, organizations are creating data services made available through application programming interfaces (APIs).

Data platforms are the centerpiece of data management. However, platforms have become more diverse as organizations seek to manage, govern, and derive value from petabytes of structured, semistructured, and unstructured data. In-memory platforms, massively parallel processing (MPP), and other technology advancements are enabling organizations to improve query performance and data availability for analytics and AI/ML.

Central to data management modernization for most organizations is cloud migration. Cloud computing arrangements vary, but the trend toward serverless computing frees organizations from having to devote most of their time to technical details such as server configuration, system and software updates, maintenance, and security. They can devote greater attention to tapping data's business potential and taking advantage of scalable and elastic cloud computing to respond to changing requirements.

Modernizing data management is not easy. As technologies and practices become established, data management becomes like a big ship that takes time to change direction. Only 8% of organizations surveyed say they are both very satisfied now with their data management and confident they can modernize technologies, services, and practices to meet evolving requirements (figure not shown).

The largest percentage is mostly satisfied with their current data management (45%), but these organizations anticipate challenges in trying to modernize. Over one-third is more pessimistic; 30% are somewhat dissatisfied with their current state of data management and not very confident in their ability to modernize, and 5% are not at all satisfied or confident they can modernize. Some say they are just getting started with data management (9%) and 3% don't know.

> *The largest percentage of respondents is mostly satisfied with their current data management (45%), but over one-third is more pessimistic.*

We will discuss data management challenges and priorities more fully later in this section. First, we examine research results about which systems, platforms, storage, cloud services, and technology standards are currently in use or planned (see Figure 7).

Not surprisingly, the most established options top the list: relational DBMS for data warehouse (68% currently using) and spreadsheets (66% currently using). Spreadsheets are ubiquitous, especially among organizations that are just starting with data management (87% of these respondents currently use them for data management, not shown). Analysts extract data from sources into spreadsheets where they then store, organize, categorize, manipulate, and analyze data.

Spreadsheets offer functions for activities such as aggregating data, performing calculations, and creating graphs. However, spreadsheets can become poor-quality data dumps when analysts lack formal processes to incorporate new data and cleanse, enrich, and prepare it. Files grow and become difficult to use and share for business insights. Data catalogs can be useful for governing data so users working with spreadsheets or other BI tools can locate and access trusted data sets.

Relational systems fit well for supporting BI reporting and analytics requirements involving mostly structured data. Transactional databases (OLTP) are also primarily based on relational technology and models (49% currently using). Application-specific databases are frequently based on relational technology. Many applications offer managed reporting and analytics functionality based on relational databases; 50% are currently using application-specific databases and 25% plan to use them.

Current and planned use of in-memory database technology is substantial. To accelerate access and improve performance, many organizations deploy in-memory databases. These have traditionally used relational models and technologies, but today columnar databases are often in-memory as are other types of NoSQL databases. As memory space has expanded, organizations are able to store entire data warehouses, analytics data platforms, data marts, and online analytical processing (OLAP) cubes in memory.

Some organizations couple in-memory OLAP with a semantic layer to optimize access to data in all locations and use the cube effectively to collect and store pre-aggregated data structures, dimensions, and measures for faster analytics. Figure 7 shows that 35% are currently using an in-memory database and 30% plan to use one.

Adoption of columnar and other NoSQL systems is growing. Growth in semi- and unstructured data and demand for complex analytics has focused attention on non-relational systems such as NoSQL data management. Although many organizations use simple data or file storage systems to hold a variety of data (55% currently use data or file storage for data management and 29% plan to use them), most organizations eventually need substantial data models and enterprise data access and management capabilities to optimize data access and deliver more complete data management. If traditional relational DBMSs are not appropriate for use cases, organizations turn to the variety of data management solutions grouped under the NoSQL umbrella.

Columnar (or column-oriented) databases, a variation on relational systems, are often included with NoSQL systems. Columnar databases store data by column rather than by row to deliver faster query performance for analytics by eliminating wasted processing from large table scans. Figure 7 shows that 36% are currently using columnar databases, which is about the same percentage (35%) as in our 2022 research. In 2022, however, only 24% planned to use a columnar database and 25% said they were not using one and had no plans to use one. This report's research notes a rise to 35% planning to use a columnar database, with only 16% saying they are not using one and have no plans to use one.

A substantial number of NoSQL systems are key-value or document databases; 31% are currently using one and 33% plan to use one. Numerous types of items could be stored as documents as required by various use cases that range from OLTP workloads to real-time, automated analytics embedded in applications. Document database solutions offer expressive query languages and multiple levels of indexing to support diverse types of queries. Key-value databases (or stores), using a simple data model of keys and values, similarly give developers flexibility to meet a variety of application and analytics requirements.

Simplicity is part of the appeal of spreadsheet-like DataFrame data structures for storing and interacting with a variety of data sourced from SQL and NoSQL files. DataFrames are used in data science, including for AI/ML projects. They are

Figure 7

Which of the following types of data management systems, platforms, storage, cloud services, and technology standards are currently in use or planned for future use at your organization?

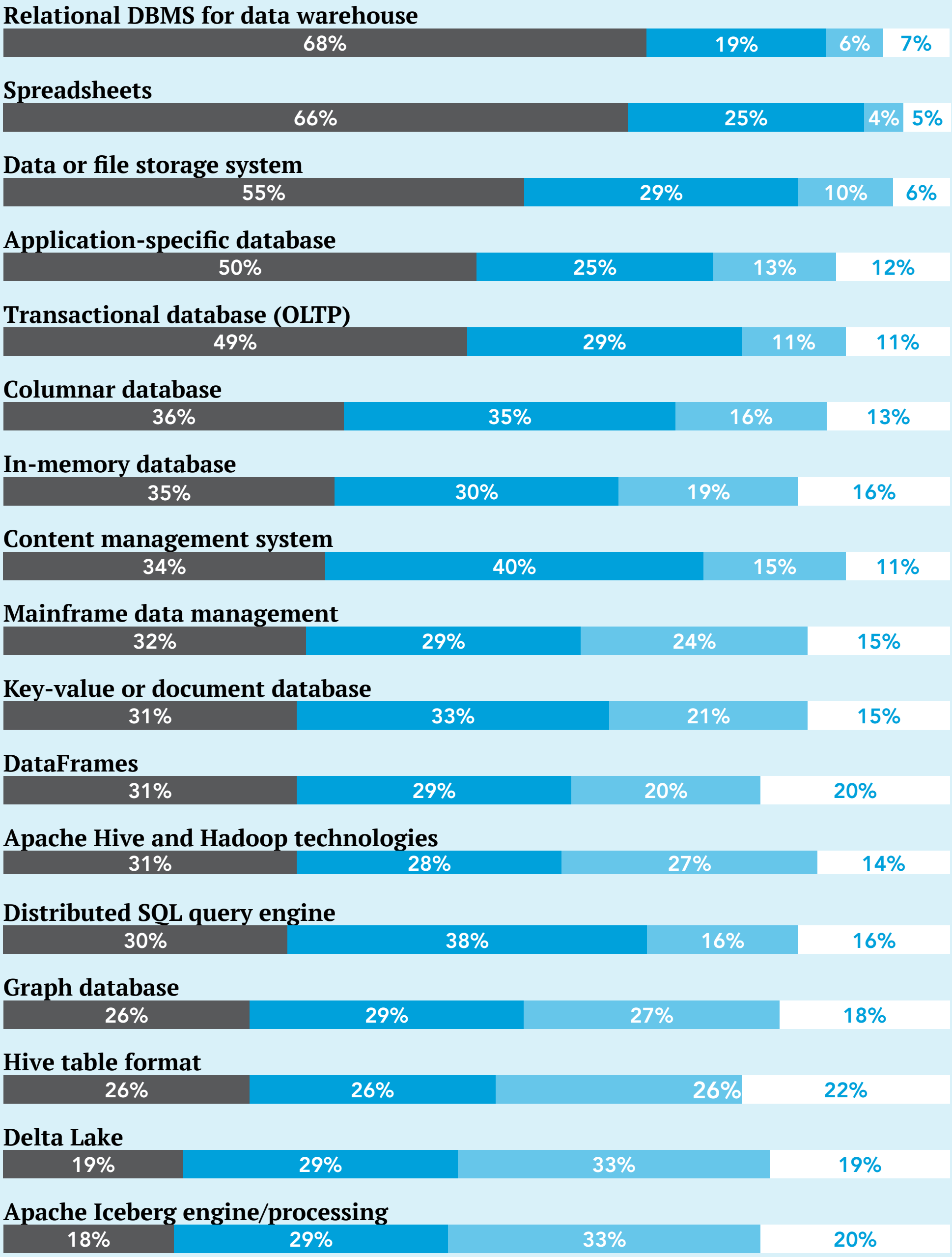
Based on answers from 171 respondents. Ordered by "currently using" responses.

■ Currently using

■ Plan to use

■ Not using; no plans to use

□ Don't know or N/A



the primary data types used in the pandas Python data analysis library and can also be used with other languages such as R and Scala. Unlike single-location spreadsheets, a DataFrame can function as a data structure shared across a distributed computing environment. In Figure 7, 31% say they are using DataFrames and 29% plan to use them.

Graph databases are NoSQL systems offering advantages over traditional relational databases for storing networked data relationships on a large and complex scale. Visual graph query languages enable users to avoid complex SQL programming and accelerate exploration and repeatable discovery of complex data relationships. Figure 7 shows that 26% are currently using a graph database and 29% plan to use one.

Cloud Data Migration and Management

Cloud data migration is a critical part of data management modernization for organizations of all sizes. Organizations want to take advantage of pay-as-you-go pricing models that offer elastic scalability. With storage and processing resource allocation in tighter sync with business needs, organizations are better prepared to respond to growth in data volume or the need for faster workload processing to accomplish seasonal or unanticipated requirements. Organizations are also taking advantage of cloud elasticity for AI/ML workloads and expansion in self-service BI dashboards and daily operational reporting.

TDWI asked research participants to choose which of a set of statements best describes their organization's status regarding cloud data management and migration. We can see that organizations are adopting a range of strategies. No one approach is dominant.

About one-quarter are focused on phased migrations to move and modernize most or all their data systems to cloud platforms (26%; figure not shown). Among larger companies with $1 billion or more in revenues, the percentage increases to 37%. With this approach, organizations migrate pieces incrementally. In each phase, organizations can test and validate related components and learn from experience with each phase how they can improve succeeding migrations. Organizations need visibility into application data dependencies to avoid disrupting use cases that are unrelated to data assets, data integration processes, and applications they are migrating.

The second-largest percentage currently uses or plans to use a lift-and-shift strategy (22%) to migrate on-premises data and workloads as-is to the cloud. For example, an organization might keep the same data warehouse model and ETL routines using the same underlying software but deploy it on a new cloud provider's platform. Organizations choose this approach to migrate more rapidly. However, it is important to balance speed benefits with taking advantage of modernization opportunities with the new cloud platforms and services.

> *With phased cloud migrations, organizations can test and validate related components and learn from experience to improve succeeding migrations.*

Hybrid data environments that comprise on-premises and cloud-based systems are common; 17% say that they plan to continue having a hybrid environment. In some cases, this is due to local data residency laws and access control requirements that make it impossible to migrate data assets. In other cases, organizations want to continue running production applications and data platforms because they are working well and have high satisfaction rates. Some organizations will keep legacy systems in place but choose a cloud-first strategy for any new systems. However, our research finds that only 11% of organizations surveyed for this report are adopting this strategy.

Data Platforms and Patterns: Current and Planned Usage

The research in Figure 7 offered a view of current and planned usage of different data management technologies and models. In Figure 8, we take a different perspective and examine the status and plans for data management platforms, repositories, and patterns for supporting BI, analytics, AI/ML, and data applications.

The largest percentage of organizations surveyed uses a data warehouse on premises (67%). Nearly half have a data warehouse in the cloud (48%) and 38% plan to implement a cloud-based data warehouse. (Note that research participants provided an answer for each type of system or pattern listed in Figure 8; we did not ask for a choice between, for example, an on-premises or cloud data warehouse.)

Across TDWI research, we have seen a strong trend toward cloud data warehouses. However, the results here show that many organizations are continuing to work with on-premises data warehouses. Many organizations indicate that they have both. Among organizations that are taking a phased migration path to the cloud, 79% are currently using a data warehouse on premises. Yet, even among these organizations, 64% also have a data warehouse in the cloud (not shown).

As we saw in our 2022 research, more organizations have a data lake in the cloud (41%) than have one on premises (30%). One-third of research participants (33%) say they plan to use a cloud data lake. TDWI defines a data lake as a design pattern and architecture optimized to capture a wide range of data types, both old and new, at scale. The purposes of a data lake are manifold and can vary among organizations.

In general, unlike data warehouses that operate with well-defined preprocessing for data transformation, formatting, and cleansing to achieve trusted, carefully structured data, a data lake lets workloads dictate how they need to categorize, cleanse, and otherwise prepare data for analytics and AI/ML. Some organizations use a data lake as a staging area for data transformation and cleansing before loading data into a data warehouse.

Unifying data warehouses and data lakes. Indeed, an important current trend is toward deploying a unified data platform such as a data lakehouse that integrates data warehouse and data lake capabilities. Figure 8 shows that 22% are currently using a unified cloud data warehouse and data lake and 46% plan to use one, which is the highest "plan to use" percentage in the figure.

A data vault is an established data modeling design pattern that some organizations view as a flexible alternative to traditional data warehouse and star schema designs. Figure 8 shows that a data vault is currently in use by 20% of organizations surveyed. About one-third plan to use it (34%). Our interview research finds that some organizations use a data vault as part of their unified data lakehouse strategy.

Cloud data lakehouses typically employ massively parallel processing (MPP) database engines that can efficiently scale computing power to fit workload requirements. Organizations could eliminate separate staging areas and choose data lakehouses as their platform for extract, load, and transform (ELT) processes, which is a variation on traditional ETL. Organizations can target data lakehouse storage for collecting and staging data; then, for the transformation, the lakehouse uses the power and scale of in-database processing.

> *Unified data platforms are adopting evolving open source file formats such as Apache Iceberg and Delta Lake storage framework to enable standardized support for all workloads.*

Unified data platforms such as data lakehouses are adopting evolving open source file formats such as Apache Iceberg and Delta Lake storage framework to enable standardized support for all workloads ranging from BI and OLAP to data science and analytics-driven applications. The standards improve support for a wider variety of programming languages and computation engines.

The standards also enable data lakehouses to standardize for ACID transactions, record-level updates, metadata standardization, schema enforcement, and additional features essential to optimizing performance for BI, analytics, and data application workloads. Referring to Figure 7, our research finds that 18% are currently using Apache Iceberg and 19% Delta Lake; 29% plan to use technologies based on each of these standards.

Nearly one-third currently use a data virtualization layer or data fabric. Figure 8 shows that 32% are currently using a data virtualization layer or fabric and 38% are planning to use one. A data virtualization layer connects to the sources to access metadata, which is then used to enable data transformations and joins that result in a new, logical data source. For users, data virtualization presents an abstraction layer, shielding them from the complexities of knowing the various source data formats and implementations to access them.

A data fabric builds on data virtualization to provide a universal and holistic approach to integrating diverse components of physically distributed data environments. This report will discuss data fabrics and related technologies in more depth later in our section on distributed data management and integration.

Data Management Modernization Priorities

As it was in 2022, data quality problems and improving the ability to fix them ranks first among data management modernization priorities for increasing user satisfaction and maximizing the value of data (44%; see Figure 9). To address data quality challenges, organizations need consistent practices supported by automated tools. Organizations should integrate data quality steps into the automation of data preparation, which 38% of respondents cite as a data management modernization priority. Data preparation and quality are not one-size-fits-all; BI reporting use cases have different requirements than data science use cases.

Tracking data lineage, noted earlier in our discussion of data governance, is essential to increasing data intelligence and solving problems faster. Figure 9 shows that this is a top priority for 35% of research participants. With data volumes growing, data lineage tracking and processes for data quality, profiling, and validation require AI/ML techniques and automation to keep pace.

Modern enterprise data catalogs offer tools that automate discovery and documentation of data lineage, including noting missing information and using AI/ML to suggest the possible lineage. Some tools provide dashboards that enable data stewards and subject matter experts (SMEs) to visualize data lineage and track the data's life cycle. Some additionally embed visibility of data quality scoring and sensitive data classification within data lineage for fuller context, impact analysis, and rapid problem identification.

Research in Figure 9 highlights the urgency to accelerate data management processes. More than one-third say it is a top priority to reduce delays in adding new data or modifying existing data (38%). Just over one-quarter of respondents (26%) are looking for alternatives to slow data movement, copying, and replication. As data volumes grow, these processes can become costly bottlenecks, which motivates organizations to evaluate new practices.

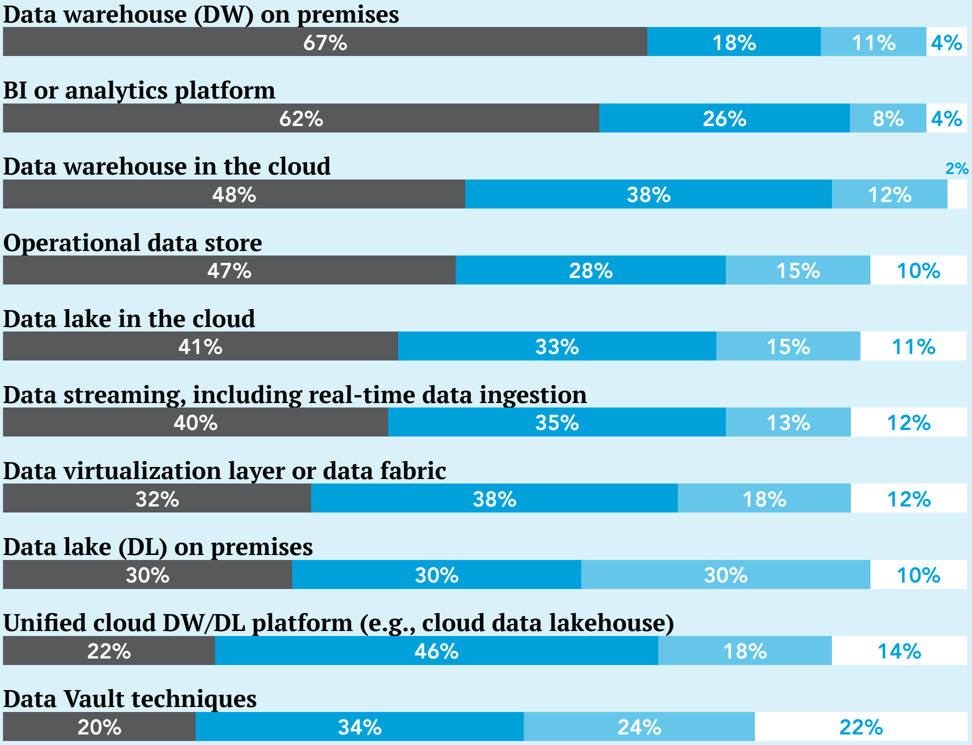
Some organizations use change data capture (CDC) technologies to capture only changes to data and data structures and provision these continuously rather than in large batch processes. Other use cases benefit from data virtualization, federation, and data fabrics to reduce data movement by connecting to and querying distributed sources, aggregating the resulting data, and providing views of it from a single point of access.

Data silo growth challenges represent a key reason behind consolidation of fragmented data into a single, unified data platform. Requiring users to locate and access data across numerous silos causes delays. Data silos make it difficult for users to gain a single view of the truth: that is, all data relevant to topics of interest. Thus, we see in Figure 9 that 34% of respondents put a priority on eliminating and consolidating disparate data silos onto a single data platform. Organizations often cannot consolidate all data silos immediately. They must prioritize which ones will most reduce data management complexity and costs and make the data environment easier to govern.

Figure 8

Which of the following types of data management platforms, repositories, or patterns are currently in use or planned for use by your organization to support BI, analytics, AI/ML, and data applications?

Based on answers from 170 respondents. Ordered by "currently using" responses.



- ■ Currently using
- ■ Plan to use
- ■ Not using; no plans to use
- □ Don't know or N/A

In the following sections we will examine data integration and distributed data management. Modernization in these areas is key to addressing data management priorities and solving challenges holistically.

Data Integration: Evolving to Meet New Challenges

Data integration technologies and practices, including data pipelines and APIs, are essential to supplying users with trusted data at the right time for reports, dashboards, real-time notifications, and analytics (including AI/ML). Organizations need to solve data integration problems that hinder decision makers from gaining actionable insights and making faster decisions. Data-driven applications with embedded analytics and automated decision capabilities are important in areas such as e-commerce, customer support, logistics, and manufacturing. These applications cannot function without seamless data integration.

The sheer number of items listed in Figure 10 makes it clear that data integration requirements encompass a range of technologies. Solutions are continuing to evolve to meet the demands of new use cases. For example, APIs have become a popular way of connecting applications for transferring data and sharing application functionality easily. APIs enable organizations to create a layer for managing connections that shields users from complexity, reducing the technical expertise required to access data. Organizations can govern data access through the layer rather than open uncontrolled data interaction. Sharing data assets and services via APIs has become an important facilitator of modern collaborative business relationships (that is, the "API economy").

Stateless Representational State Transfer (REST) APIs offer the most established approach and are popular for their simplicity and flexibility. However, an alternative query language for APIs, GraphQL, is growing in popularity among application developers and data analysts. GraphQL enables users to examine complete descriptions of data available through APIs and query the data more efficiently. GraphQL enables single-request aggregation of data from multiple sources or APIs. In Figure 10, we see that APIs are currently used by the largest percentage of organizations surveyed (59%) and a significant percentage plan to use APIs (28%).

> *Data connectivity needs to support user demands for easier setup, expanded data discovery and findability, and significantly reduced latency.*

The second-largest percentage say they are using data connectors and connectivity tools (46% currently using and 32% planning to use). This is a broad category of tools that are important for delivering data to applications either from original sources or intermediate data platforms. Data connectivity needs to support user demands for easier setup, expanded data discovery and findability, and significantly reduced latency through real-time reporting and analytics. It is important for organizations to monitor data connectivity to improve standardization; legacy data connectivity is often characterized by numerous point-to-point integrations that are conflicting, redundant, and difficult to update. Monitoring is also important to prevent data governance exposures.

Data transformation and pipelines are commonly used. ETL and ELT are next highest, used by

45% and 44% of organizations respectively. With ELT, organizations build data pipelines to load (or replicate) data into a target repository before data transformation, thereby saving data movement to a separate staging area. The research shows that traditional ETL, however, is still important and is the appropriate choice for certain workloads.

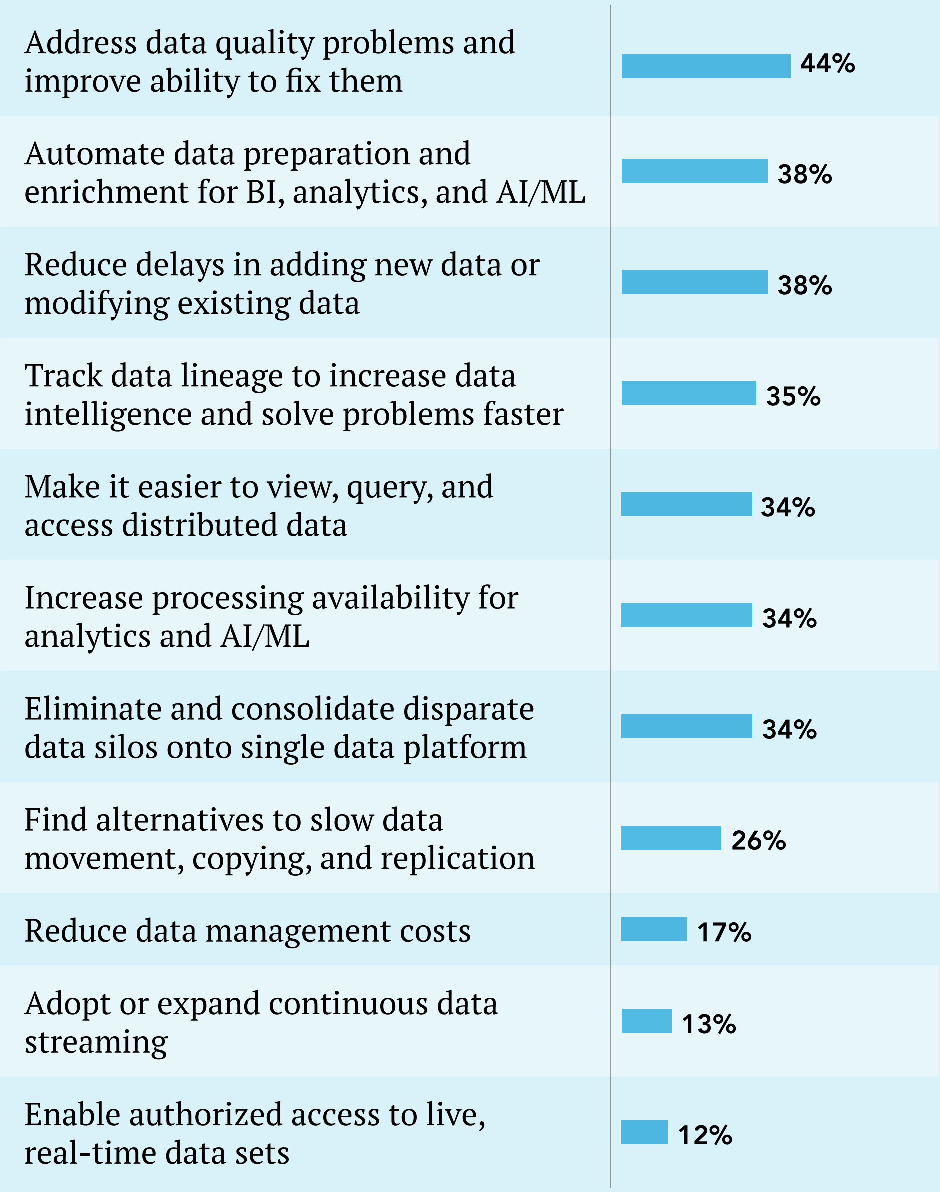
Data pipelines, which often contain ETL or ELT functionality, are in use by 43% of organizations and 35% plan to use them. Some data pipelines simply connect to sources and stream raw, even real-time, data into a data lake or data lakehouse. This is primarily for use by data scientists performing predictive analytics and running AI/ML programs. Other pipelines provide fuller data profiling, quality, and validation steps as needed by workloads.

Data preparation and preprocessing have a significant role in data integration. The data pipeline steps mentioned above (plus others such as data enrichment) fall under the umbrella

Figure 9

Overall, which of the following are your organization's top current data management modernization priorities for increasing user satisfaction and maximizing the value of data?

Based on answers from 167 respondents, who were asked to select at least their top three objectives.



of data preparation and preprocessing. Data preparation processes focus on determining what the data is and improving its quality and completeness. Organizations use data catalogs in data preparation steps to standardize how different users define and structure the data.

The life cycle of preparation and preprocessing steps starts with collecting and consolidating data and continues through transformation and enrichment to make data useful for purposes such as reporting, dashboards, OLAP analytics, and AI/ML. Self-service data preparation is a critical trend for empowering nontechnical users as well as data scientists to be less reliant on IT developers and analysts.

Two out of five respondents (40%) in Figure 10 say their organizations are using data preparation or preprocessing tools and 34% are planning to use them. Data preparation and preprocessing steps can be time-consuming and involve considerable manual work, which makes tools that automate steps and integrate preparation with data pipelines highly valuable. Some organizations are employing data fabric preparation and delivery layers to simplify data exploration and transformation in heavily distributed data environments.

Options for accelerating access to timely data are popular. Organizations are currently using or are planning to use several technologies that enable them to bypass lengthy data integration and preparation steps to allow users to access fresher, more frequently updated data. More than one-third are using CDC technologies (36%), and 39% plan to use them to capture changes to data continuously rather than force users to wait for batch updates to historical data.

Nearly as many are currently using data streaming (e.g., real-time) technologies (33%) and more are planning to use them (40%). Data streaming solutions offer continuous flows through pipelines that gather and process data as sources generate it. Organizations apply real-time analytics and AI/ML to track situations, understand patterns, and answer complex business questions.

> *Over a third (36%) are currently using a data catalog or metadata management system and 47% plan to use one, which is the highest "plan to use" percentage for any item in Figure 10.*

Sizeable percentages use data intelligence technologies. Data intelligence consists of knowledge about diverse data built on metadata resources contained in a repository such as a data catalog. As shown in Figure 10, 36% are currently using a data catalog or metadata management system and 47% plan to use one, which is the highest "plan to use" percentage for any item in the figure.

Along with accurate data definitions and metadata, shared data intelligence systems will manage and include data profiles, information about the data's location, taxonomy and classification documents, calculations, and data lineage information about sources and data ownership. Significant percentages of organizations surveyed are using solutions based on open source standards. Nearly one-third (30%) are currently using Apache Atlas, a data governance and metadata framework; 25% are planning to use it. About one-quarter are either currently using (23%) or planning to use (25%) Metastore, an open source Apache Hive metadata repository typically used with data lakes.

Figure 10

Which of the following types of data integration, data catalog, and data intelligence technologies are currently in use or planned for use at your organization?

Based on answers from 167 respondents. Ordered by "currently using" responses.

- ■ Currently using
- ■ Plan to use
- ■ Not using; no plans to use
- □ Don't know or N/A

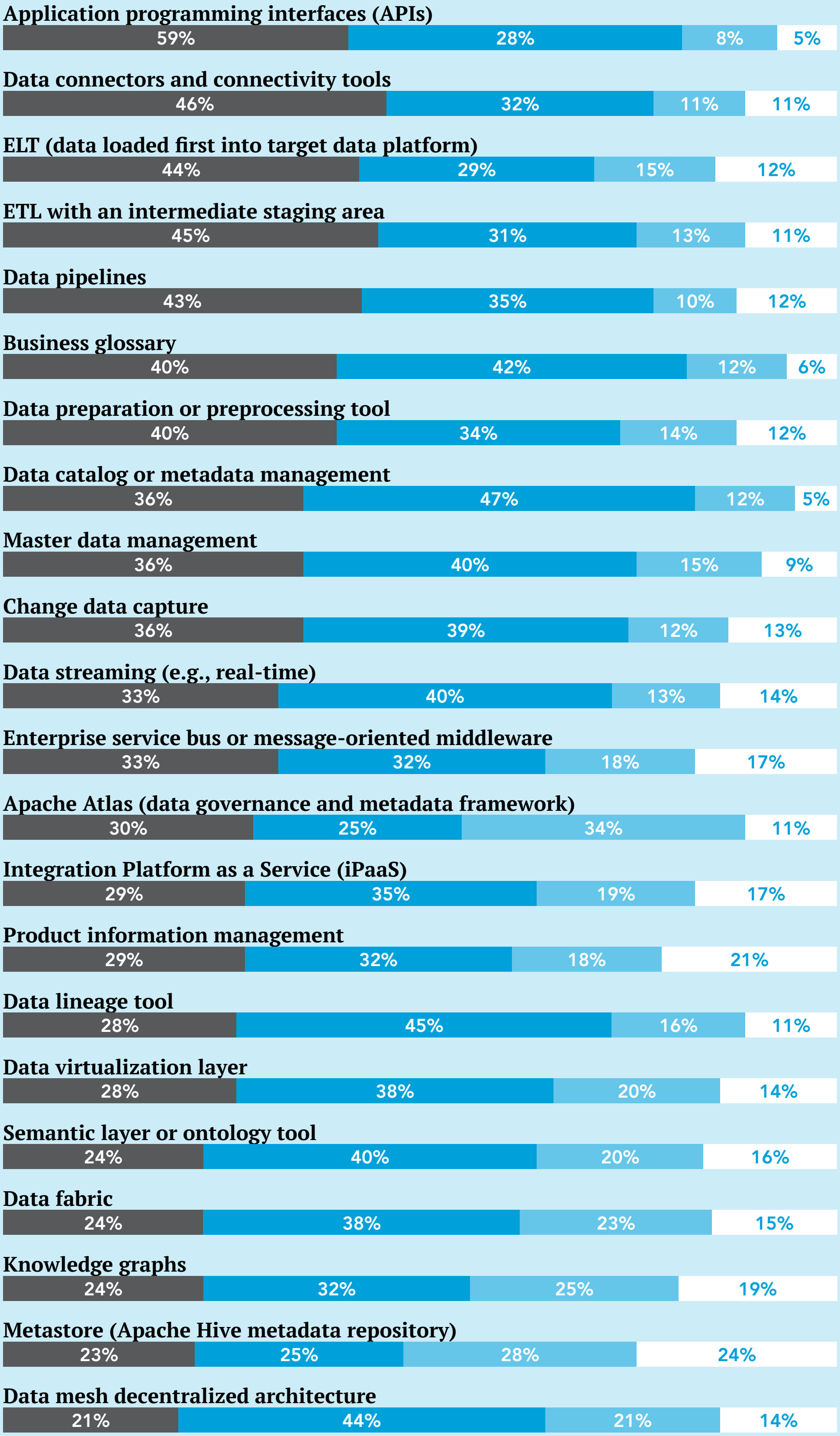


Figure 10 also shows strong interest in data lineage tools, which are incorporated as capabilities in some data catalogs and data integration suites or are offered as standalone data intelligence tools. TDWI finds that 28% are using data lineage tools now and 45% plan to use them. We will discuss key goals with data catalogs, business glossaries, and metadata management—and levels of satisfaction in achieving them—later.

Also contributing to data intelligence are business glossaries (40% currently use and 42% plan to use them) and master data management (MDM), which 36% currently use and 40% plan to. Business glossaries collect, organize, and coordinate business terms and definitions to provide clarity and data context across organizations. Business glossaries may also contain business rules, business policies, data sharing agreements, and other business assets to be used in conjunction with technical data assets. Modern data intelligence solutions merge business glossaries and data catalogs together.

MDM systems document and coordinate master data, which is higher-level data about products, suppliers, customers, and other business entities. MDM enables organizations to establish "golden" master data definitions that are consistent, valid, and accurate across applications and services. Many organizations today are establishing MDM systems that integrate master data from different domains (most commonly customer and product information) to reduce management overhead and enable cross-domain data access and analytics.

Figure 10 shows that 29% are currently using a product information management (PIM) system and more (32%) are planning to use one. In research for this report not shown here, TDWI finds that 29% are using a customer information management (CIM) system to capture, manage, and govern customer data while 20% are using a multidomain MDM system for this purpose.

Objectives for Improving Data Integration

With the preceding discussion of current and planned data integration systems in mind, we now turn attention to improvement priorities. Improving data quality, accuracy, and completeness is the top data integration objective (51%; figure for this section not shown), which aligns with what we reported for data management modernization. Organizations should create processes for determining data quality and use automated tools to monitor changes in the data. AI-infused automation in data intelligence tools is enabling organizations to move past the limits of manual data curation.

Reducing the amount of data movement, replication, and copying is an important objective for 34% of organizations surveyed. As data volumes rise, these processes often cause data latency and therefore slower time to insight. Automating data movement to save cloud costs and optimize processing is a top objective for 27% of organizations. As we noted earlier, heavy data movement, replication, and copying also raises data governance concerns because of potential exposure of sensitive data. Cost also rises. A primary driver behind interest in data virtualization and data fabrics is reducing complexity and costs due to data movement, replication, and copying.

Cloud migration of data integration tools and systems is a priority. One-third of organizations (33%) are seeking to migrate data integration, pipelines, and data intelligence to the cloud. It is important to gather knowledge about existing on-premises data integration to inform cloud migration teams about dependencies and best strategies for migration. Knowledge about current ETL performance will help organizations eliminate unnecessary workloads—a priority of 25% of research participants—as they shift data integration processes to the cloud. Organizations can use migration as an opportunity to switch to ELT, which for 20% is a top objective.

In some cases, using phases to migrate data integration processes in smaller units is the best approach, but in others it may make sense to migrate all related processes and subprocesses together. With either strategy, using automated tools will increase consistency and accelerate testing of newly migrated processes.

Finally, 32% cite the importance of reducing costs and improving monitoring of cost drivers. Large organizations often have thousands of ETL, data pipeline, data streaming, and data preparation processes. Modernization should include monitoring to improve visibility into the value of data integration processes and whether resources need to be reallocated to optimize the most important processes and eliminate those that are no longer needed.

Performance optimization focuses on reducing costs through higher efficiency. Modern, automated tools enable performance optimization without user intervention. Optimization should ensure that workloads have the resources they need and show expected rates of consumption. Once you have granular visibility and proper controls in place, you can use tool capabilities in cloud data platforms to identify inefficient spending.

Cost optimization and FinOps practices can be valuable. Maintaining continuous awareness of cloud data resource consumption is key to keeping spending in line with overall financial plans. As part of cost optimization, some organizations establish cloud financial operations, or FinOps. Inspired by DevOps practices, FinOps frameworks facilitate collaboration among stakeholders in engineering, finance, technology, and business domains regarding spending decisions. Monitoring visibility into real-time data about costs, use, and allocation is essential to FinOps. Tools in the marketplace are evolving to provide granular visibility and use AI/ML to generate recommendations.

Data Catalogs: Current Satisfaction and Where to Expand

Data catalogs, business glossaries, and metadata management systems share capabilities. They can play complementary roles, with business glossaries specializing in business definitions. In the marketplace, data catalogs, business glossaries and metadata management systems are generally converging into single, all-in-one data intelligence solutions.

> *With organizations supporting more diverse users and workloads, a data catalog provides a necessary foundation for both defensive and offensive data governance.*

A centralized metadata repository such as a data catalog helps organizations reduce confusion about the data’s quality and consistency and enhances data curation and governance. A complete and accurate data catalog includes what the data means and represents, the data's origins, where to find it, who is responsible for it, and information to help the user quickly understand if it is relevant to a user's objectives. With organizations supporting more diverse users and workloads than ever before, a data catalog provides a necessary foundation for both defensive and offensive data governance and plays an important role in making it easier to leverage and share high-value, trustworthy, and relevant data and data intelligence with other users.

To understand where organizations are realizing benefits and where they need to improve, TDWI research examined levels of satisfaction with data catalogs, business glossaries, and metadata management systems (see Figure 11). All the objectives listed in the figure are important; we discussed some of them earlier in the context of defensive and offensive data governance. Here, for brevity, we will highlight those with the highest levels of satisfaction.

- **Making it easier for users to search for and find data.** At the point of data use, users need the ability to access the data catalog through an intuitive interface to discover relevant, available data and learn about its consistency, lineage, age, and governance constraints. Organizations show three times greater satisfaction today than when we asked about this objective in 2021; 32% are very satisfied compared to just 10% in 2021. About the same percentages are and were somewhat satisfied (36% in this report versus 35% in 2021).

- **Improving data governance and regulatory compliance.** As noted earlier, data catalogs are valuable for their role in managing metadata to document accurate data inventories, which is necessary for the regulatory compliance aspects of data governance. Data catalogs can also manage data governance and security rules that drive constraints in data pipelines and applications. Overall, 63% are satisfied with their data catalog, business glossary, or other metadata management system for improving data governance and regulatory compliance, with 22% very satisfied. The percentages for inventorying data assets and documenting objects available to users are similar; 21% are very satisfied and 42% are somewhat satisfied.

- **Improving data pipelines, transformation, and preparation.** In Figure 11, just over three in five respondents (61%) are satisfied with their data catalog, business glossary, or metadata management system for this objective, with 20% very satisfied and 29% unsatisfied. This also shows improvement over our 2021 research findings, which showed 11% were very satisfied, 31% were somewhat satisfied, and 28% were unsatisfied. Closer integration of data catalogs with data curation processes saves users' time in locating relevant data and accelerates development of trusted data products.

Automation in data catalogs enables faster identification of data quality, consistency, and completeness issues. Figure 11 shows that 60% are satisfied with these data intelligence technologies for improving these areas. During data pipeline development and transformation and preparation processes, some data catalogs can automatically expose which data sets are available, trusted, and governed.

Where organizations appear to be encountering the most difficulty is in establishing a single, location-transparent resource for metadata. Two in five respondents (40%) say their organizations are unsatisfied, the largest percentage in Figure 11. When organizations have numerous departmental databases or multiple data platforms, data warehouses, and data lakes, they can suddenly be drowning in disparate metadata that is difficult to align.

Modern data catalogs have AI-infused capabilities for crawling distributed data platforms for metadata to coordinate data definitions and meaning across platforms. When organizations establish an enterprise data catalog, it can serve as a single, location-transparent resource regardless of the location of the data or type of data architecture.

Distributed Data: Challenges and Opportunities

Distributed data environments mostly just happen. As organizations grow, add subsidiaries or divisions, and reorganize to compete in new markets, business data requirements generate new data platforms and applications. Users need more than what monolithic enterprise data systems provide. Then, technology evolution, especially for analytics and data science, generates new systems that grow alongside existing legacy systems. Rapid migration to the cloud typically adds to the amount of distributed data, which spreads across a hybrid of on-premises and multiple cloud-based data platforms. As a result, data silos proliferate and become difficult to access, integrate, manage, and govern.

> *Distributed data environments hold considerable potential if organizations can bridge data silos and manage the environments holistically.*

Although there is no single way to solve these problems, distributed data environments hold considerable potential if organizations can bridge data silos and manage the environments holistically. However, when growth is piecemeal rather than according to an enterprise strategy, data integration and management becomes dauntingly complex. In Figure 12, we can see that organizations are currently using or plan to use a variety of strategies to address distributed data challenges and maximize the potential of all the diverse data.

Organizations are consolidating data silos. Consolidating data from silos into a single data platform is the current strategy for the largest percentage of organizations surveyed (47%). Reducing the number of siloed data systems can make data management and governance easier and potentially reduce total cost of ownership (TCO). Organizations no longer need expert administrators for each platform or to carry licenses or subscriptions for more systems than they need.

Consolidating to a modern cloud data platform enables organizations to take advantage of elastic scalability. Capabilities such as automatic scaling (“auto-scaling”) in leading cloud platforms can add or subtract computational resources in response to changes in workloads without intervention from users or administrators. Modern platforms use automation and rolling technology updates to continuously optimize performance even as workloads increase in number and complexity.

The research in Figure 12 shows the importance of data intelligence tools and systems such as a data catalog and semantic layer; 32% are currently using these to address distributed data challenges and 52% plan to use them. Earlier, this report discussed the value of location-transparent enterprise data catalogs. Additional semantic layer technologies play an important role in ensuring the quality and consistency of business representations of distributed data for BI reporting, calculations, and multidimensional modeling.

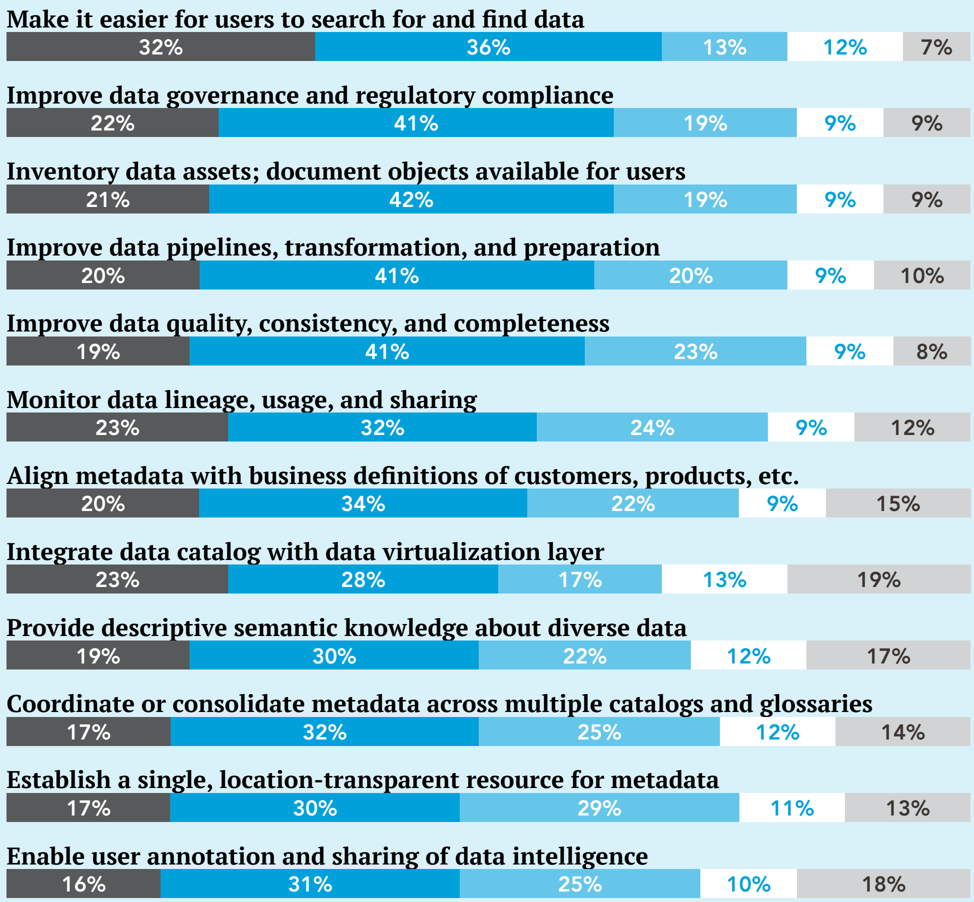
Some organizations are using enterprise knowledge graphs and ontologies to improve discovery of how distributed data sets relate to each other and to definitions of higher-level entities such as people, places, and things that might be contained in MDM. In Figure 12, multidomain MDM is currently in use by 28% of organizations surveyed and 39% plan to use this solution. One-quarter of respondents (25%) say their organizations are using knowledge graphs to visualize and analyze distributed data relationships and 38% plan to use them.

Accelerating analytics and AI/ML depends on distributed data management and integration. Nearly one-third are currently unifying analytics to work across domains, query engines, and storage repositories and 42% plan to do this. Access to all relevant data from across distributed systems is critical to analytics and AI/ML. A major challenge is to hide the technical complexity of accessing each system so that nontechnical users and data scientists can move faster to gain single views. Semantic data integration that brings

Figure 11

How satisfied is your organization with its data catalog, business glossary, or other metadata management system for achieving the following objectives?

Based on answers from 166 respondents. Order by combined "very satisfied" and "somewhat satisfied" responses.



■ Very satisfied

■ Somewhat satisfied

■ Somewhat unsatisfied

□ Not at all satisfied

■ Don't know or NA

together data catalogs and data virtualization layers is an important strategy. Figure 12 shows that 30% are currently using a data virtualization layer and 40% plan to use one.

A similar percentage of respondents are currently using a distributed SQL engine to query data in multiple sources (28%) or planning to use one (42%). Often based on open source Presto, a distributed SQL engine can use a single SQL query to retrieve and interact with large volumes of data at multiple sources using models and systems ranging from traditional relational DBMSs to NoSQL systems and data lakes.

Data Fabric and Data Mesh Architectures

Figure 12 shows that over one-quarter of organizations are developing a data fabric to view and query distributed data (27%). A data fabric provides a location-transparent layer, usually built with data virtualization, to make it easier and faster to access distributed data and manage and govern it holistically. The fabric should be extensible and modular to ensure less dependence on central IT.

Instead of a piecemeal approach, the goal of a data fabric is to have an organized and orchestrated strategy for onboarding new data across the distributed environment. It enables organizations to set up policies and standards for governed data pipelines and data curation processes and for deploying computational resources to maximize benefits.

> *Instead of a piecemeal approach, the goal of a data fabric is to have an organized and orchestrated strategy for a distributed environment.*

A key goal of a data fabric is to take advantage of opportunities to automate tedious and repetitive data engineering tasks. Solutions are also focusing on using AI/ML to generate recommendations for locating relevant data and streamlining preparation tasks. Organizations are also automating application of data governance constraints within data fabrics.

Data intelligence is critical to data fabrics and data meshes. Data catalogs, active metadata management, and semantic data integration technologies play important roles in providing shared knowledge about distributed data for faster, easier, and more consistent access, data preparation, and governance. Data intelligence is also important to creating a decentralized data mesh architecture, which 21% of organizations surveyed are currently using.

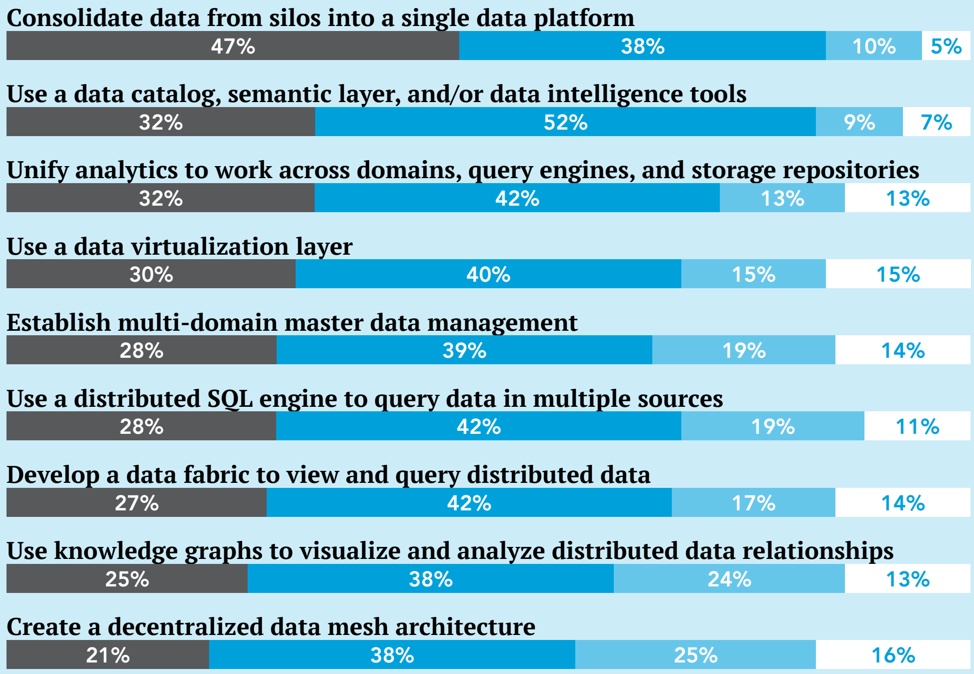
The data mesh has emerged as a leading strategy for enhancing the value of distributed data and avoiding the costs, delays, and management headaches of copying and moving massive amounts of data. Most data mesh strategies focus on strengthening distributed data ownership and governance within domains: that is, where users with a direct relationship to the business priorities build, manage, and share data products across the organization. A data mesh can complement a data fabric and take advantage of both data virtualization and open data lakehouses to enable distributed data governance from centralized policies and increase the value of data.

A data mesh supports “data as a product” development. It encourages teams working in domains to treat data sets as products that have “customers” who could be business users, data scientists, data engineers, or external users of monetized services, potentially made available through a data marketplace or exchange. The

Figure 12

To address distributed data challenges, which of the following strategies for unifying management and governance and integrating data access are currently in use or planned to be used by your organization?

Based on answers from 169 respondents. Ordered by "currently using" responses.



■ Currently using
■ Plan to use
■ Not using; no plans to use
□ Don't know or N/A

notion is that organizations need to take the same care in developing and operationalizing data sets and data products that they do in delivering traditional products to customers. Depending on the use case, customers may need data sets enriched with metadata and additional capabilities. Data products increase in value when people can trust them for decisions and collaborate to achieve business objectives.

DataOps for Planning and Scalability

With development and production of data products, data pipelines, and analytics, organizations need frameworks to provide a structure for planning, orchestration, and scalability. DataOps builds on agile and DevOps methodologies to provide a structure that has similarities to software development life cycles. Organizations use DataOps and data observability capabilities to monitor and manage progress toward operationalization. They can spot bottlenecks and quality problems that are not only slowing progress but are also affecting the organization's ability to achieve business objectives.

We asked research participants if their organizations are currently using or planning to use DataOps practices and whether they have had experience in using DataOps or related methods (see Figure 13). Only 15% say they are currently using some or all DataOps practices successfully. This is somewhat low, but we are still in the early stages of maturity in use of DataOps.

> *Only 15% say they are currently using some or all DataOps practices successfully. This is somewhat low, but we are still in the early stages of maturity.*

In the first stages, organizations may adopt only some elements of the full framework. As the figure shows, more organizations have experience in using established agile and DevOps methods (26%). A significant percentage are planning to use DataOps but have not yet begun (21%). Smaller percentages say they are not having success with DataOps or have no experience or interest in it.

To gauge what issues are driving organizations to use a framework such as DataOps, we asked research participants what their most challenging issues are in improving delivery of scalable and repeatable value from data, code, and analytics models. Here are the top issues for respondents (figure not shown):

- **Improving communication and collaboration among stakeholders (45%).** Collaboration is critical to ensuring quality, sorting out dependencies, and eliminating redundancy. Improving collaboration and communication are central objectives for establishing continuous feedback loops in the DataOps framework.

- **Having flexibility to adjust when requirements change (39%).** Projects often lose momentum and focus when users change requirements. Agile and/or DataOps principles for shorter iterations can help prepare project teams to handle new requirements.

- **Increasing automation throughout data integration and preparation (37%).** Using a framework plus modern data observability tools can help organizations understand which phases of integration and preparation processes need automated tools to reduce manual work and increase consistency.

- **Ensuring data governance throughout data integration and preparation (36%).** Data governance is a key focus of DataOps. Visibility through data observability monitoring tools can enable organizations to understand the complete health of data environments and be able to address both defensive and offensive data governance issues quickly.

- **Streamlining delivery of data sets, models, analytics, and other artifacts (31%).** Organizations cannot streamline delivery when quality is a problem. Agile, DevOps, and DataOps all help organizations spot quality problems sooner, before they have a negative impact on downstream data products, analytics models, and other artifacts. Data observability tools can improve end-to-end visibility to identify and resolve problems hindering the speed of delivery into production.

Data Objectives for Optimizing Analytics and AI/ML

With demand rising for accelerated development and deployment of advanced analytics and AI/ML, it is critical to address data access, preparation, and management problems that prevent projects from achieving success. To close this report's research, we asked participants to identify their highest priority objectives for optimizing analytics and AI/ML workloads and increasing business value (see Figure 14).

The priority shared by most organizations surveyed highlights the importance of data democratization: making it easier for users to discover and access all types of data (45%). Organizations are seeking solutions and practices that will empower users to uncover insights drawn from exploration of relationships between sets of structured, semistructured, and unstructured data. Nearly one-quarter prioritize making it easier and faster to discover data relationships (23%). Organizations will need to take measures to improve data availability and support for self-service ad hoc querying.

The second most common objective is to make it easier for non-coders to search, access, and prepare data (42%). This again highlights the trend toward data democratization; organizations do not want to unnecessarily limit deeper data interaction to only technically skilled programmers and developers. One-quarter of research participants see enabling self-service data preparation (e.g., data loading, blending, and transformation) as one of their keys to optimizing analytics and AI/ML workloads. Along with making data discovery and access easier, organizations need tools that empower nontechnical users, business analysts, and data scientists to prepare data as they see fit with less dependence on IT.

Trusting predictive models and the results of AI/ML depends on appropriate data curation. A significant percentage of respondents are looking for improvement in curating diverse data sets to ensure unbiased results (28%). Organizations need AI model governance tools that provide visibility into training data selection and curation to avoid “black box” models (that are not transparent). AutoML support is a focus, with some research participants looking to reduce delays in preparing data processes; 18% say it is a priority to set up accelerators and predefined data processes for autoML.

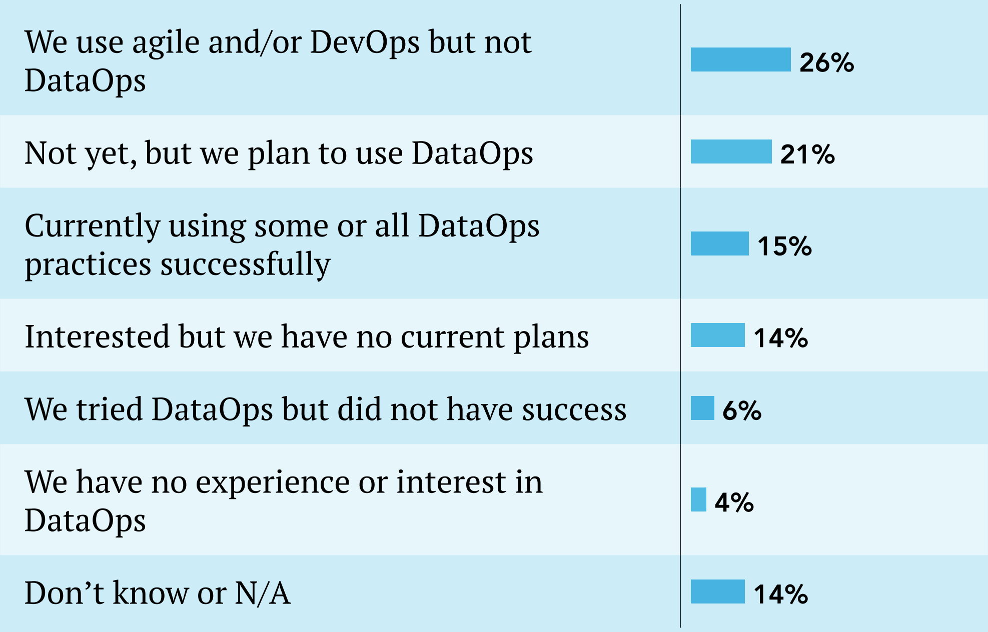
Using automation to accelerate data insights is a shared priority. Nearly one-third of research participants prioritize automating discovery of actionable data insights (31%). Technology advances across data integration, data intelligence, and front-end data applications are aligning to enable automated discovery. In Figure 14, 22% say that they regard it as a top priority to augment BI, KPIs, and metrics with AI-driven recommendations. Surfacing timely, intelligent recommendations to augment users’ dashboards can accelerate informed business decisions. To enhance predictive understanding of future directions in sales, operational expenditures, and business profitability, an even larger percentage say that enriching business forecasting with AI/ML insights is important (30%).

> *One in five organizations surveyed say that enabling use of LLMs and generative AI is a top objective (20%).*

Generative AI is a rising objective. One in five organizations surveyed say that enabling use of LLMs and generative AI is a top objective (20%). Across industries, there is clearly tremendous interest today in the transformational capabilities of generative AI—a subset of AI/ML that offers techniques for generating new content in a variety of forms, including but not limited to text, code, images, and speech. Generative AI requires robust data platforms to support LLMs and similar ML models at the core of generative AI. Organizations will need to revise defensive and offensive data governance rules and policies to guide growth in generative AI.

Figure 13

Is your organization currently using or planning to use DataOps practices to optimize productivity and support growth in data pipelines and analytics and AI/ML workloads? Have you had experience in using DataOps or related methods?



Based on answers from 163 respondents.

Recommendations

Here are 10 best practices for achieving scalable, agile, and comprehensive data management and governance. This report has discussed how data democratization, data-driven applications, and advancement with analytics and AI/ML are major drivers behind modernization. Our research shows that these 10 best practices are key for meeting demands generated by these drivers.

Harness modern technologies and practices to improve data quality. The top priority across much of our research is improving trust in data quality, accuracy, and completeness. With new data sources and types of workloads, data quality will never be a fully solved problem. Organizations need to take advantage of data catalogs, data preparation tools, and capabilities in data platforms and virtual layers to align data curation processes with user requirements. AI-infused automation can help organizations improve data quality in massive, high-velocity data.

Enhance data management and governance agility. When markets, economic conditions, and competition are changing quickly, business leaders and data scientists need agility so they can focus on solving business challenges. Good practices for data governance can fall by the wayside, but this eventually creates trouble, including higher costs. Our research shows dissatisfaction with data management and governance support for strategic decision-making and new business development. Bring business leadership and stakeholders together to find the right balance between agility and governance. Use the elasticity of cloud data platforms to handle changing requirements. Automating constraints can play a beneficial role in making data governance less obtrusive and easier to update.

Address analytics and AI/ML pain points. Our research shows that solving data management, integration, preparation, and governance challenges is key to moving forward with augmented analytics, including generative AI. Reducing time and costs associated with data collection, preparation, and provisioning for feature engineering is a priority. Organizations should evaluate automated tools and examine processes to rationalize those that are redundant. Ease of use and automation are advancing in solutions to make it easier for non-coders to discover and prepare data.

Establish data stewardship. Many of the challenges faced by organizations in our research call for improved data stewardship. Data stewards have expertise in the data and can oversee both defensive and offensive data governance. They can mentor new users to follow rules and policies and apply best practices for selecting data sets that are fit for purpose. As noted, data stewards are usually busy with "day jobs" in IT as data analysts and managers or have business roles as subject matter experts in business domains. Thus, automated tools and automated processes in data catalogs are critical to reducing hands-on monitoring and manual work.

Improve data intelligence and maximize the value of data catalogs. Data intelligence, centered in a data catalog, is critical to both defensive and offensive data governance. Data catalogs and other metadata management tools help keep track of and classify sensitive data for protecting PII. They also aid in discovering data quality and consistency issues and streamline solving them. Modern data catalogs offer AI-infused automation to handle complex, high-volume data and feature easier-to-use interfaces for less technical users.

Improve self-service data integration, transformation, and pipeline development. Organizations would like to reduce the time users spend on data preparation and pipeline development. There is strong interest in empowering users to do more of their own work rather than depend entirely on IT. In this report, we find that demand for self-service capabilities continues to grow. However, organizations should increase the role of shared resources in data integration, transformation, and pipelines—such as data catalogs and the computational power of cloud data platforms—to improve speed, scale, and quality, and reduce redundancy.

> *To improve self-service satisfaction, organizations should increase the role of shared resources—such as data catalogs and the computational power of cloud data platforms*

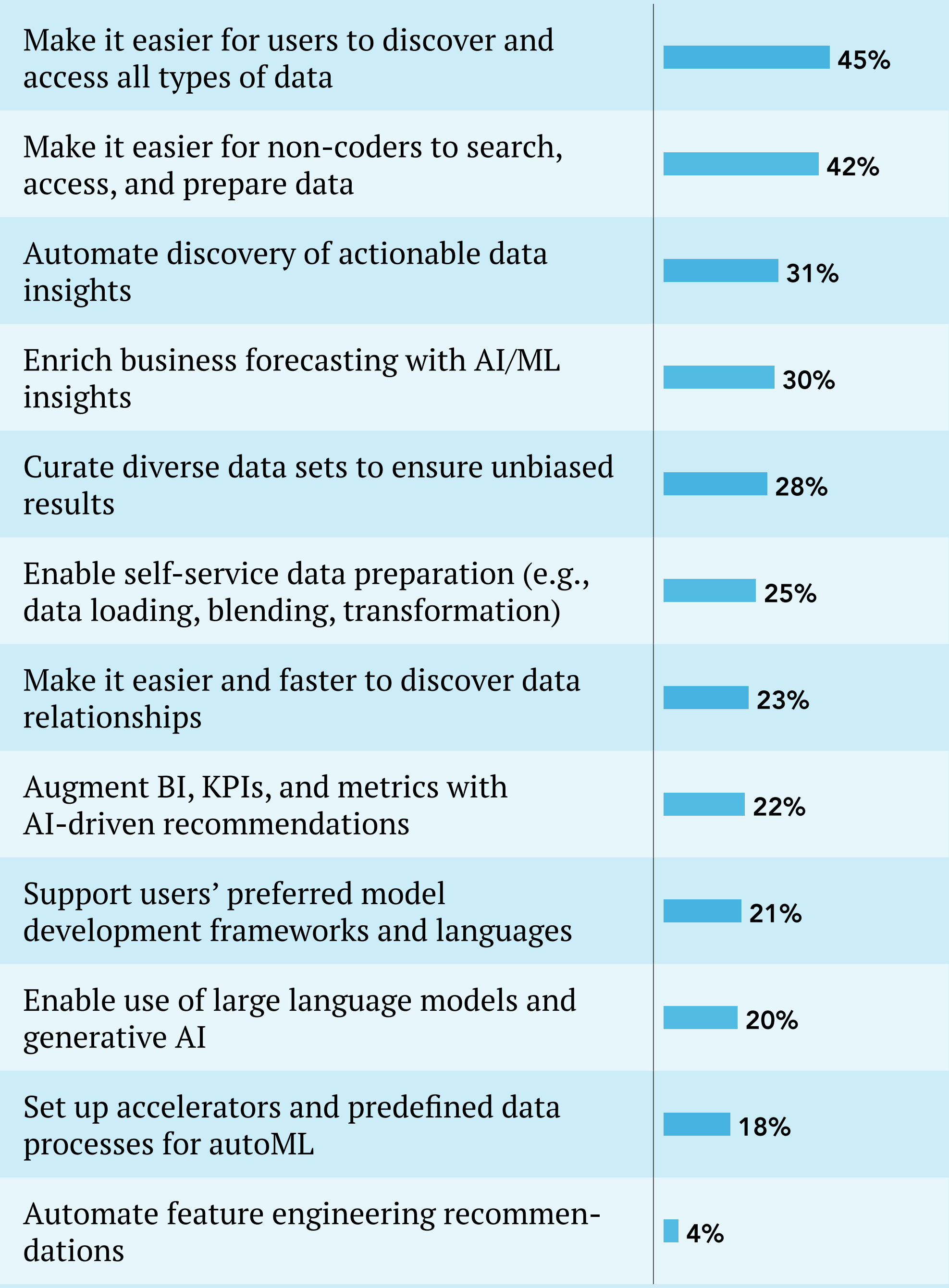
Examine alternatives to heavy data movement, replication, and copying. Our research finds that reducing these activities is one of the top objectives among organizations surveyed. Automation, monitoring, and scheduling help identify and solve bottlenecks. Consolidation to a unified data platform can reduce some needs for data movement, such as between data staging areas and target platforms. In other cases, organizations can avoid movement by setting up a data virtualization layer and developing a data fabric architecture.

Evaluate a data fabric architecture for distributed data. Research in this report shows considerable interest in data fabrics to improve data access, governance, and management of distributed data. Growth in hybrid multicloud data environments is accelerating interest as organizations evaluate how to establish a virtual, location-independent layer above multiple data platforms. Data fabrics rely on semantic richness based on metadata and additional data intelligence, often managed by a data catalog. Organizations should evaluate current

Figure 14

Which of the following objectives are your highest priority to address for your organization to optimize analytics and AI/ML workloads and increase their business value?

Based on answers from 159 respondents, who were asked to select at least their top three objectives.



components to determine if they can support a data fabric and modernize where needed.

Consider a data mesh architecture. With business domains across organizations increasingly focused on developing various types of data products, they are seeking empowerment and reduced dependence on centralized data platforms and enterprise IT. A data mesh champions decentralization to enable domain teams closest to business objectives. As they empower domains, organizations also need to enable cross-domain projects, enterprise data governance, cost management, and efficient resource sharing. Evaluate a data mesh architecture as a potential framework for decentralization and self-service to enhance business agility. Use federation and automation of constraints to standardize data governance.

Evaluate DataOps for better planning and scalability. Unplanned increases in data pipelines, data preparation, and development of data products can generate costly problems without a management framework. Our research finds growing interest in DataOps. Improving communication and collaboration among stakeholders is a challenge faced by many organizations and is a key goal of DataOps. Evaluate data observability monitoring as part of DataOps to improve end-to-end visibility into development and the business impact of data management and governance problems.

Hitachi Vantara

Hitachi Vantara is part of Hitachi Group. From transportation and energy to automotive systems—the group focuses on sectors that make an incredible impact on our future. Understanding how data enables mission-critical digital and industrial environments is what we do.

The Pentaho platform enables insights at scale into data by operationalizing data management with automation accelerating data discovery, onboarding, and tiering, as well as data pipeline orchestration and analytics. This improves data quality and trust by discovering, identifying, categorizing structured and unstructured data into meaningful terms defined by customers, and standardizing them through business glossaries. This helps organizations meet their data governance gathering and usage requirements. In addition, the analytics tools provide users with observability of their pipelines and can provide dashboards to ensure data quality-driven outcomes.

The Pentaho platform portfolio includes:

- **Pentaho Data Integration & Analytics:** end-to-end IT and OT data integration and analytics for any data type

- **Pentaho Data Catalog:** provides trusted data by automating discovery for self-service analytics, compliance, and quality across an enterprise-wide data fabric

- **Pentaho Data Storage Optimizer:** tiers data between file systems such as Hadoop and S3 Object Storage seamlessly for data management and cost savings

Learn more on hitachivantara.com

TDWI Research provides research and advice for data professionals worldwide. TDWI Research focuses exclusively on data management and analytics issues and teams up with industry thought leaders and practitioners to deliver both broad and deep understanding of the business and technical challenges surrounding the deployment and use of data management and analytics solutions. TDWI Research offers in-depth research reports, commentary, inquiry services, and topical conferences as well as strategic planning services to user and vendor organizations.

tdwi | TRANSFORMING DATA WITH INTELLIGENCE™

A Division of 1105 Media
6300 Canoga Avenue, Suite 1150
Woodland Hills, CA 91367

E info@tdwi.org

tdwi.org