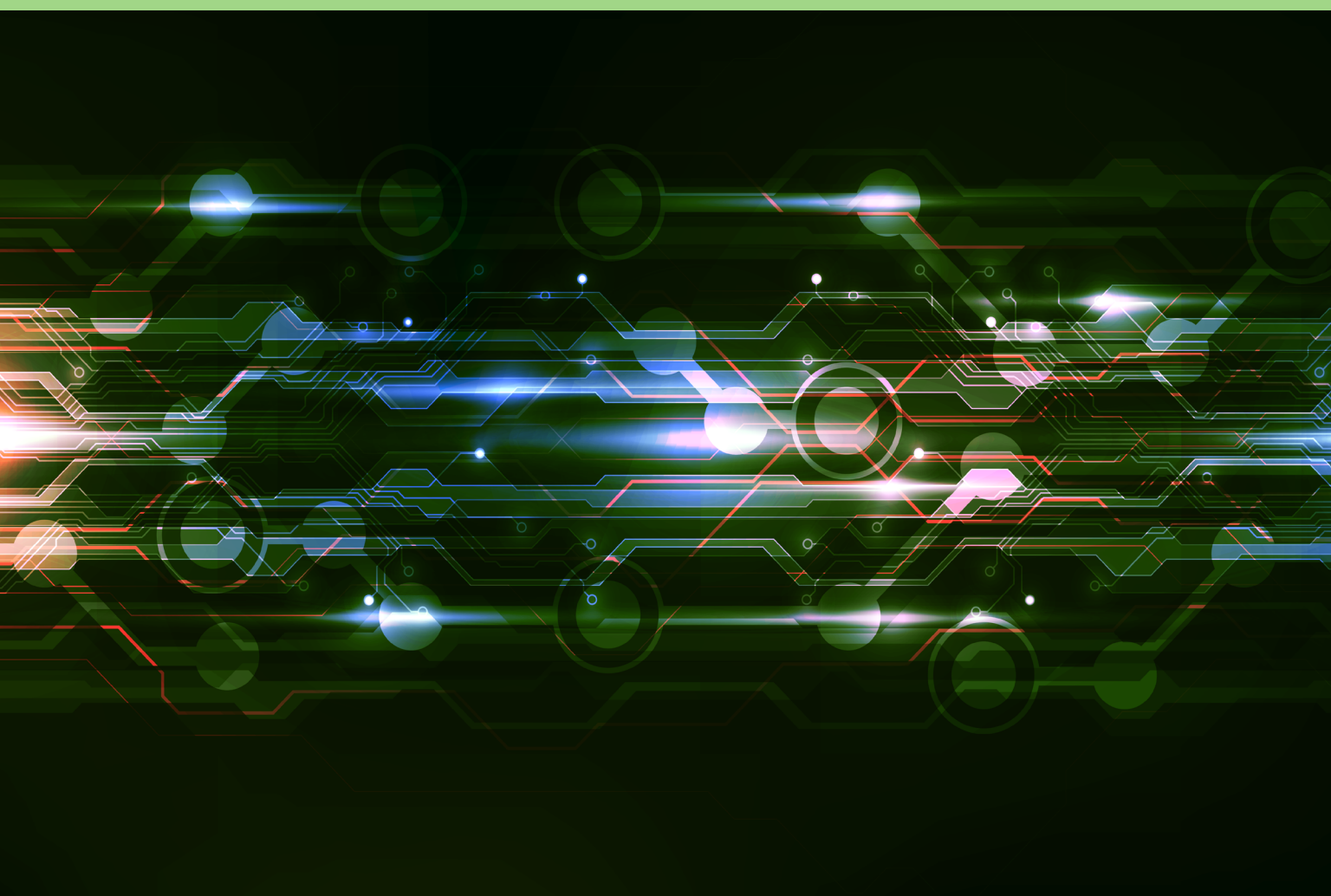
Co-sponsored by:

MongoDB.

BEST PRACTICES REPORT

Q3 2022

Maximizing Business Value with Data Platforms, Data Integration, and Data Management



By David Stodder

tdwi TRANSFORMING DATA WITH INTELLIGENCE™

BEST PRACTICES REPORT

Q3 2022

Maximizing Business Value with Data Platforms, Data Integration, and Data Management

By David Stodder

Table of Contents

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.

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, Alteryx, Denodo, MongoDB, SAP, and Snowflake sponsored the research and writing of this report.

About the TDWI Best Practices Reports Series

This series is designed to educate technical and business professionals about new business intelligence 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 is supplemented by surveys of business intelligence professionals.

To support the program, TDWI seeks vendors that collectively wish to evangelize a new approach to solving business intelligence 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 Research analysts Fern Halper (fhalper@tdwi.org), David Stodder (dstodder@tdwi.org), James Kobielus (jkobielus@tdwi.org), and Markum Reed (mreed@tdwi.org).

About TDWI Research

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

Research Methodology and Demographics

Report purpose. Data democratization, digital transformation, cloud migration, distributed data environments, and the development of data-rich, AI-infused applications are major trends driving modernization of data platforms, data integration, and data management. This TDWI Best Practices Report examines modernization priorities. It discusses how organizations can overcome challenges (such as outmoded legacy practices, outdated technologies, and data silos) to maximize the value of data assets.

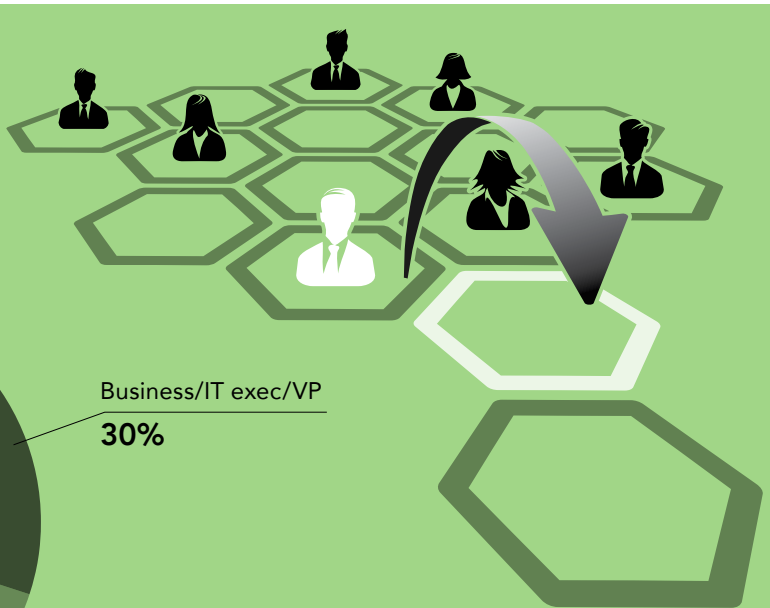
Survey methodology. In late June and July 2022, TDWI sent invitations via email to business and IT professionals in our database, asking them to participate in an internet-based survey. The invitation was also posted online and in publications from TDWI and other firms. The survey collected responses from 381 respondents, with 310 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.

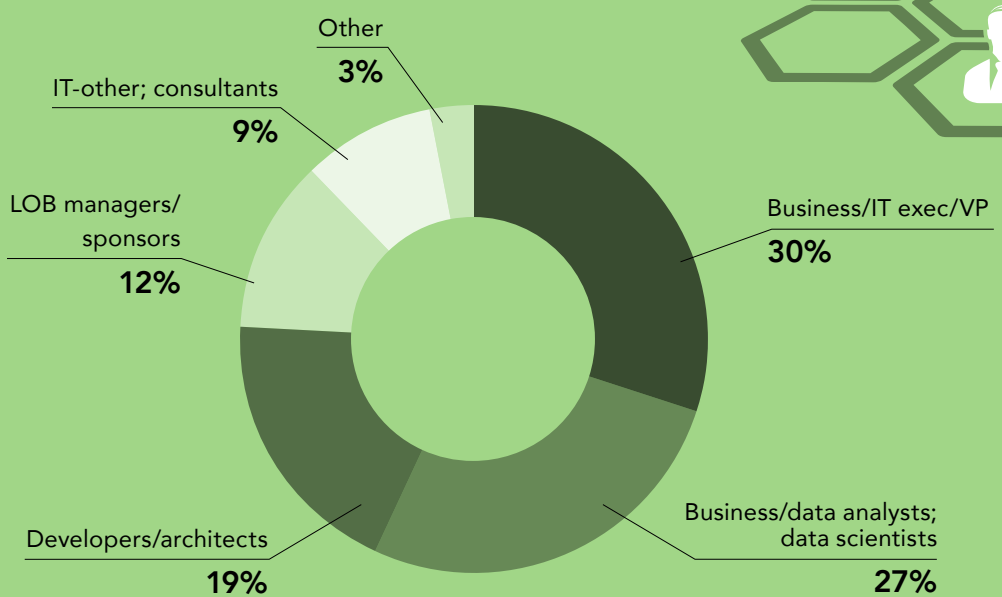
Research methods. In addition to the survey, TDWI conducted interviews with business and IT executives and managers, application developers, and data management and analytics experts. TDWI also received briefings from vendors that offer products and services related to the topics addressed in this report.

Survey demographics. Nearly one-third of respondents are business or IT executives and VPs (30%). The second-largest group consists of business or data analysts and data scientists (27%). Third largest is developers and data, application, or enterprise architects (19%). Line-of-business (LOB) managers and business sponsors account for 12% of the respondent population. Other IT staff, consultants, and other titles account for 12% of the total.

Industries varied considerably. Software/internet is the largest (19%), followed by financial services and healthcare (10% each), consulting and professional services (7%), education and government (6% each), insurance (5%), and manufacturing (non-computers) (5%). Just over half of respondents reside in the U.S. (52%), with South Asia (primarily India and Pakistan) (20%), Asia and Pacific Islands (8%), Canada (5%), and other regions following. Respondents come from enterprises of all sizes.

Position

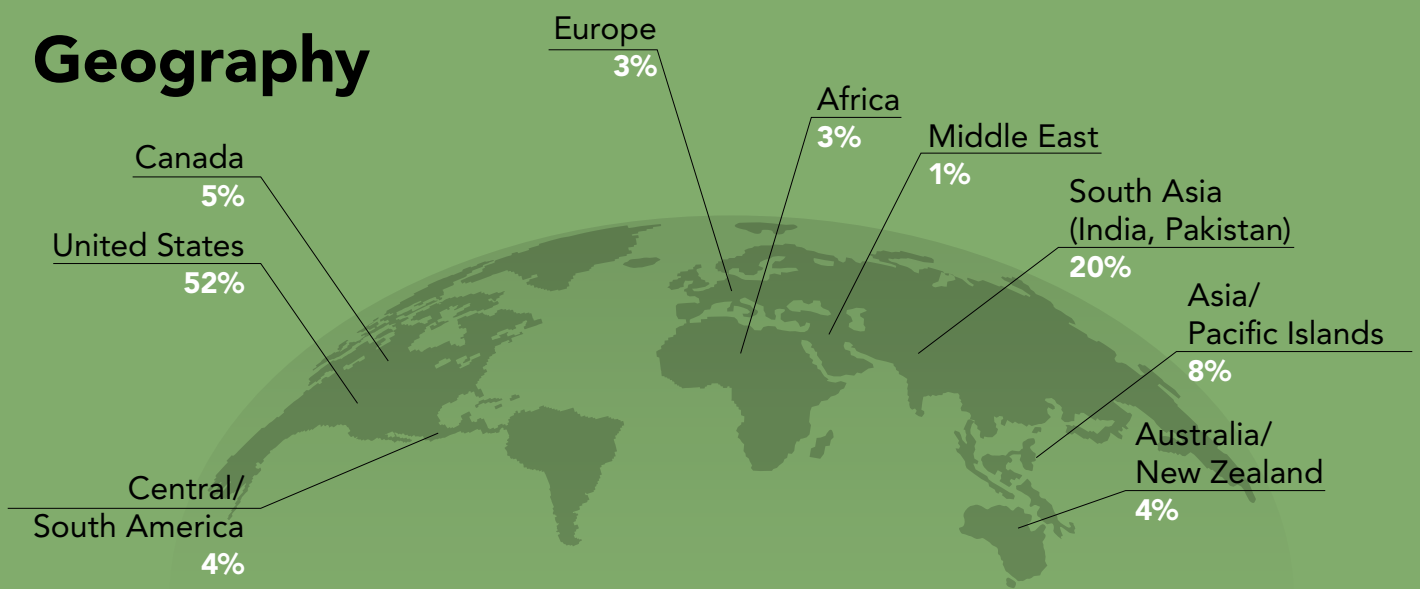




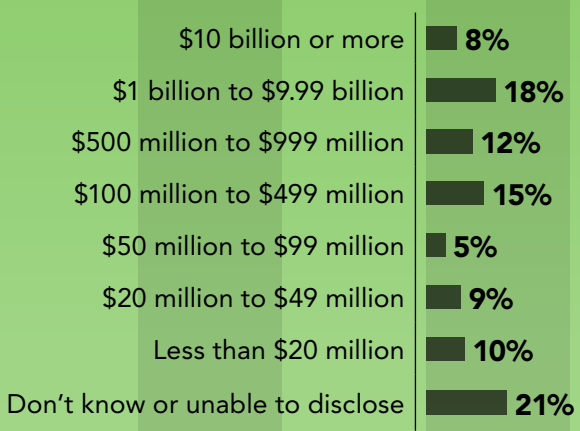
Industry

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

Geography



Company Size by Revenue



Based on 310 respondents.

Executive Summary

Maximizing the value of data platforms with data integration and data management capabilities is essential for organizations to become data- and analytics-driven. Increasingly—but not entirely—based in the cloud, these systems and services are the engine of modern applications as well as smarter processes for higher business efficiency and innovation. Business imperatives such as digital transformation, personalized customer engagement and marketing, agile and resilient supply chains and manufacturing, and faster response to unexpected situations are driving interest in modernization.

> *Organizations need faster, in-context data insights, increasingly through development of AI/ML. This report reveals where organizations are facing challenges and how they can overcome them.*

In this TDWI Best Practices Report, organizations show strong interest in modernization; 54% are actively modernizing now and 29% plan to do so in the near future. Legacy data systems rooted in fixed configurations and one-size-fits-all solutions for business intelligence (BI) reporting and analytics are inadequate. Too many organizations are mired in data silos and custom code for data integration and management.

Organizations need faster, in-context data insights, increasingly through development of artificial intelligence and machine learning (AI/ML) models and algorithms to drive modern decision-making and automated actions inside applications and business processes. Users are demanding self-service BI and analytics. They need data integration to empower them to answer business questions faster and collaborate on decisions more efficiently. This report reveals where organizations are facing the greatest challenges and how they can overcome them.

Research indicates the need for tighter integration between two traditionally separate worlds: one devoted to analytics generation and the data integration, platforms, and management that support analytics, and the other devoted to mission-critical business applications and processes essential to operations across enterprises.

Digital transformation and cloud migration are major drivers of change. Innovative data applications depend on continuous data flows and intelligent insights based on high-volume and highly varied data. In this report, organizations show interest in gaining the benefits of a range of data management systems and platforms to maximize the value of all their data and use new, active approaches to provisioning the right data at the right time to more users.

Distributed data scenarios create challenges, including hybrid multicloud environments that combine on-premises systems and services based on multiple cloud provider platforms. Organizations report concerns about data quality; fixing data quality issues is a top priority. However, distributed data environments often feature numerous data silos that contribute to poor data quality as well as incompleteness and inconsistency problems. Some organizations are consolidating data silos into central, unified data platforms in the cloud. They show interest in more tightly integrated data warehouses and data lakes. Others are using data virtualization layers to enable trusted data views drawn from multiple sources.

Building resources of knowledge about an enterprise's data is vital for supporting different types of workloads. Data intelligence is also critical to meeting data governance priorities for protecting sensitive data and locating and accessing trusted data sets faster. Semantic layers are essential to building knowledge resources about data and aligning business representations with the data amid change. This report discusses the importance of enterprise data catalogs for meeting goals with semantic layers to make data governance more comprehensive and enable faster, more trusted data insights.

The report concludes with a discussion of how technologies and practices are coming together to create unified data environments, including with emerging data mesh and data fabric frameworks. It discusses how modernization makes this unity flexible rather than restrictive. The balance enables organizations to empower teams to maximize the value of enterprise data assets. We close the report with 10 best practices recommendations for success.

Modernization: Keeping Pace with Business Change

To compete in today's fast-changing business environment, organizations need data platforms, data integration, and data management systems and services that are agile, empower decision-makers at all levels with trusted data, and enable organizations to move past the constraints of legacy data systems and applications. Although most organizations continue to have some on-premises systems, cloud computing is changing the dimensions of what is possible.

> *Modern cloud-based data platforms and integration services provide elastic scalability to handle accelerating growth.*

Modern cloud-based data platforms offer elastic scalability to handle growth in data and demands for faster processing to support a wider range of workloads. The workload spectrum runs from BI dashboards to advanced data science using artificial intelligence (AI) techniques such as machine learning (ML). Distributed data integration technologies and frameworks are also modernizing through cloud services that enable organizations to view, query, manage, and govern data wherever it resides, including both on premises and on multiple cloud platforms.

For many organizations, the priority is to migrate to the cloud as soon as possible. Some organizations attempt "lift-and-shift" migrations where they keep workflows, models, and other properties unchanged when they move from on premises to the cloud. The main risk with such migrations is that organizations can miss opportunities to modernize capabilities. They may not take full advantage of cloud computing benefits such as cost elasticity, scalable performance, and reduced maintenance. Hidden complexities, including undocumented dependencies, can slow lift-and-shift cloud migrations.

TDWI finds that many organizations prefer business-driven, phased migrations. These focus each phase on data and processes that are relevant to particular lines of business (LOBs), departments, applications, analytics projects, or data platforms. Lift-and-shift strategies have their place, but phased migrations generally aim to solve business problems rather than just move data. They limit disruptions so unrelated

workflows continue and data remains available for users across the enterprise.

Organizations of all sizes have ambitions to be data-informed and data-driven. They want to empower users to interact with data easily and share trusted insights within the business. This is the essence of data democratization, which is about enabling more types of users, including external partners and customers, to tap the power of integrated data. Through self-service BI and analytics, users receive the insights at the right time to carry out their responsibilities and improve business outcomes.

To satisfy democratized users, organizations are employing scalable cloud data platforms with built-in data integration and data management capabilities or a set of independent data integration and management technologies and services. Increasingly, modern cloud services enable users to tap the power of AI/ML within applications and cloud services without having to become full-time data scientists.

Cutting-edge applications maximize the value of data assets. Many organizations are developing (or are contracting with third parties to develop) data-driven applications. These are business and operational applications designed to maximize the value of data assets by driving real-time, automated decisions in response to customer behavior, events, pattern detection, and more.

Application developers increasingly use containers (and open source Kubernetes for orchestrating containers) to give organizations the flexibility to integrate data-rich component services for vertical needs. These could include services for data visualization, data observability, analytics, AI/ML models, and selective data feeds. In many cases, cloud-based marketplaces and exchanges enable organizations to select prebuilt services.

Today's data applications require bridging the traditional gap that has BI, analytics, and supporting data platforms on one side and operational business applications for managing and processing transactions, e-commerce, customer relationships, inventory management, and other vertical applications on the other. With modern data applications, developers embed and operationalize analytics, data science, and visualizations. This requires data architectures that support continuous flows between application processes and the data systems supporting analytics and AI/ML models. Data management must increasingly support around-the-clock, concurrent interaction by more users than typical traditional BI and data warehousing. This includes nontechnical LOB users who are focused on a business process, not on gaining the expertise needed to write complex queries.

Drivers for Data Modernization

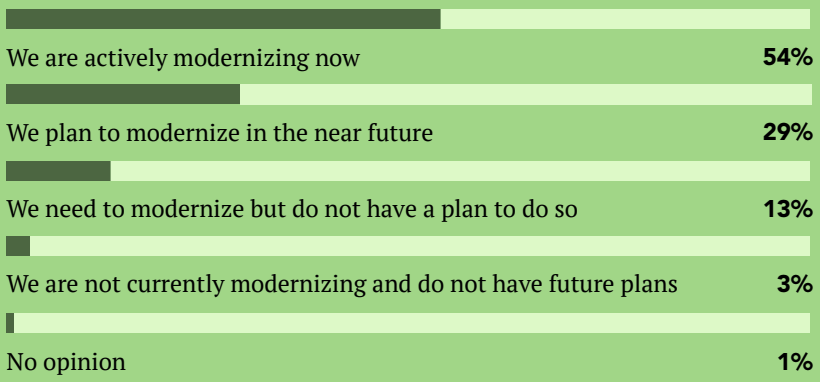
These trends put pressure on legacy data platforms as well as independent data integration and management systems. With legacy systems, IT organizations have had to configure data systems to fit fixed, on-premises storage and processing constraints and to serve limited populations of skilled analysts, developers, and power users.

To open the research survey for this Best Practices Report, TDWI asked participants whether their organizations are currently modernizing data platforms, data integration, and data management by investing in new technologies, cloud-native services, skill sets, and processes, or whether they plan to do so in the near future (see Figure 1). The

Figure 1

Is your organization currently modernizing its data platforms, data integration, and data management by investing in new technologies, cloud-native services, skill sets, and processes or does it plan to in the near future?

Based on answers from 381 respondents.



results show a strong appetite for modernization. More than half say their organizations are actively modernizing now (54%) and a significant percentage plan to do so in the near future (29%).

What does *modernization* mean? Essentially, it means updating data platforms, including their capabilities for data integration and data management, to take advantage of the latest technologies and cloud services. Modernization enables organizations to grow their use of data, innovate, increase efficiency, and improve user satisfaction. It also means updating data governance practices and adjusting the organization's culture to support data-driven objectives. This report will examine current experiences and practices with technologies and discuss strategies for modernization.

Enterprises are guided by multiple business imperatives fueling modernization. To learn more, we asked participants to rank the most important business drivers of current or planned modernization initiatives (Figure 2).

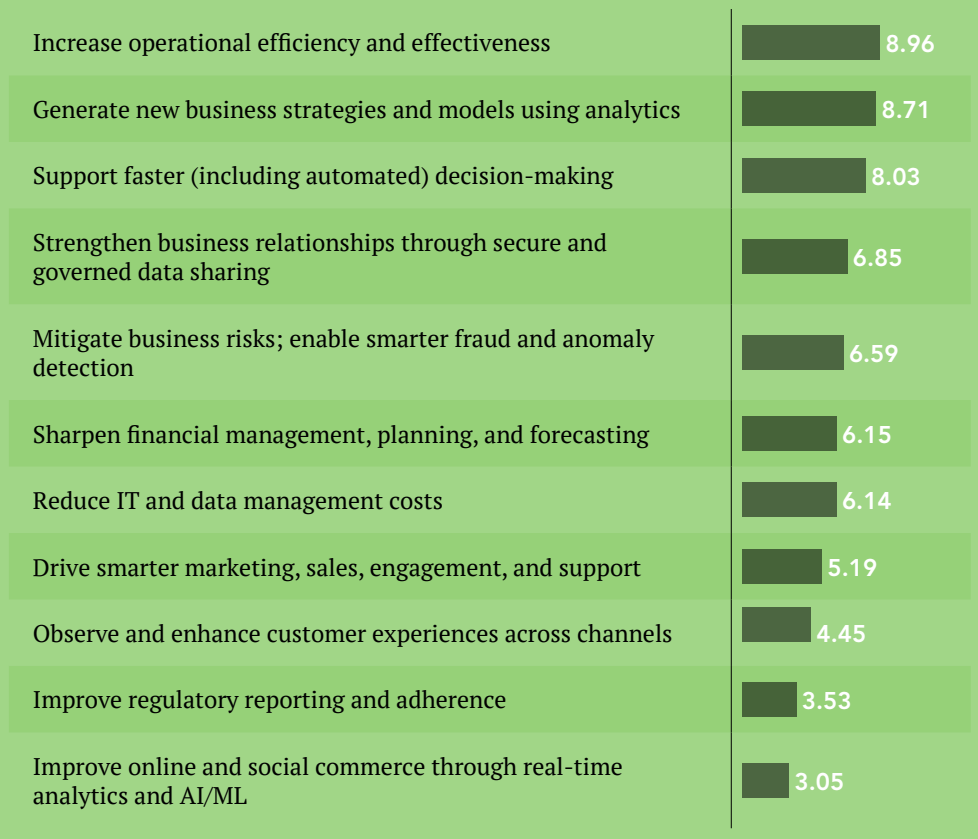
The top-ranked driver is to increase efficiency and effectiveness in business operations (25% of participants made it their top pick). Always a leading driver, business operations need data insights that reveal how managers can improve processes, reduce duplication and inconsistency, and ease bottlenecks, thereby reducing costs and increasing customer satisfaction.

> *Increasing operational efficiency and effectiveness is the top-ranked driver behind modernization initiatives; generating new business strategies and models ranks second.*

Data infrastructure is critical to delivering data to the right people at the right time; this increases efficiency and effectiveness and supports analytics for deeper insights that lead to operational innovation. However, today this requirement extends beyond the use of BI and analytics systems to include data-driven, mission-critical operational applications. Many of these applications need a data infrastructure that supports real-time data feeds and analytics at the point of human decisions or automated actions.

Figure 2

What are the most important business drivers behind current or planned initiatives by your organization to modernize data platforms, management, and integration?



Based on answers from 381 respondents, who were asked to rank the options. Ordered by weighted average.

This report discusses these types of applications and their requirements in a later section.

Every objective listed in Figure 2 is important, but the top-ranked business drivers reveal the range of current concerns. After increasing operational efficiency and effectiveness, these four drivers fill out the top five:

- **Generating new business strategies and models using analytics** (for 41% of respondents this was their first choice). Many industries are undergoing significant, sometimes unexpected, changes due to factors such as COVID-19 pandemic-related changes in customer behavior, challenges to supply chain resilience, difficulty hiring skilled workers, and economic stresses such as inflation. Inflexible data infrastructures make it difficult for organizations to explore new data, develop innovative analytics models, and use data insights to update application and business processes.

- **Support faster (including automated) decision-making.** Organizations surveyed show strong interest in modernizing data infrastructure to eliminate latency in making data-informed decisions, including automating decisions inside applications. To accomplish this objective, organizations are tightening integration between analytics and applications.

- **Strengthen business relationships through secure, governed data sharing.** Access to shared data and analytics continues to rise in importance for business-to-business

collaboration and for collaboration between internal and subsidiaries' teams. Today, data sharing through participation in data marketplaces (either internal or external) and exchanges facilitates easier collaboration. Forward-looking data infrastructures need to support expanded data sharing.

- **Mitigate business risks; enable smarter fraud and anomaly detection.** The pandemic and supply chain problems have made resilience a top priority, but organizations have long sought to apply predictive models and other advanced analytics to discover better ways of reducing business risk and detecting fraud and abuse. The pace of business change, including establishing new digital channels for customer or partner engagement, creates new risks. Channels such as social networks make fraud and abuse perpetrators harder to identify. To respond, organizations need modern data infrastructures that enable faster access to new data types collected from multiple sources and the identification of data relationships. The architecture must support continuous modeling processes for risk analytics and automated fraud detection.

Data Strategy and Digital Transformation

Digital transformation is a major trend. Organizations are migrating formerly manual, offline applications and business processes to more agile and automated environments that run on software and maximize the value of data. Today, digital transformation typically includes cloud migration and the use of software-as-a-service (SaaS) solutions. Digital transformation often generates big data—high-volume, high-velocity, and highly varied data. The new data and the potentially easier paths to collecting and integrating the data increase opportunities for deploying analytics and AI/ML to help achieve the business drivers listed in Figure 2.

> *To gain the benefits of digital transformation, it is critical to align the transformation with your organization's data strategy.*

To gain the benefits of digital transformation, it is critical to align the transformation with your organization's data strategy. A *data strategy* is an overarching plan that unifies decisions involving data technology and data process modernization.

Once an organization develops an overall data strategy, it needs to keep the strategy up to date. Business and IT leaders must ensure that the strategy is forward-looking so their organization is prepared for both the opportunities and the challenges posed by digital transformation.

Data strategy best practices include:

- Focusing modernization of data management and integration on maximizing the value of data assets (i.e., doing more than merely amassing data in storage; ensuring that data integration is delivering value)
- Improving user data literacy and addressing change management issues that arise with digital transformation
- Solving data flow problems that obstruct fast, confident decision-making (both by humans and within AI-infused applications)

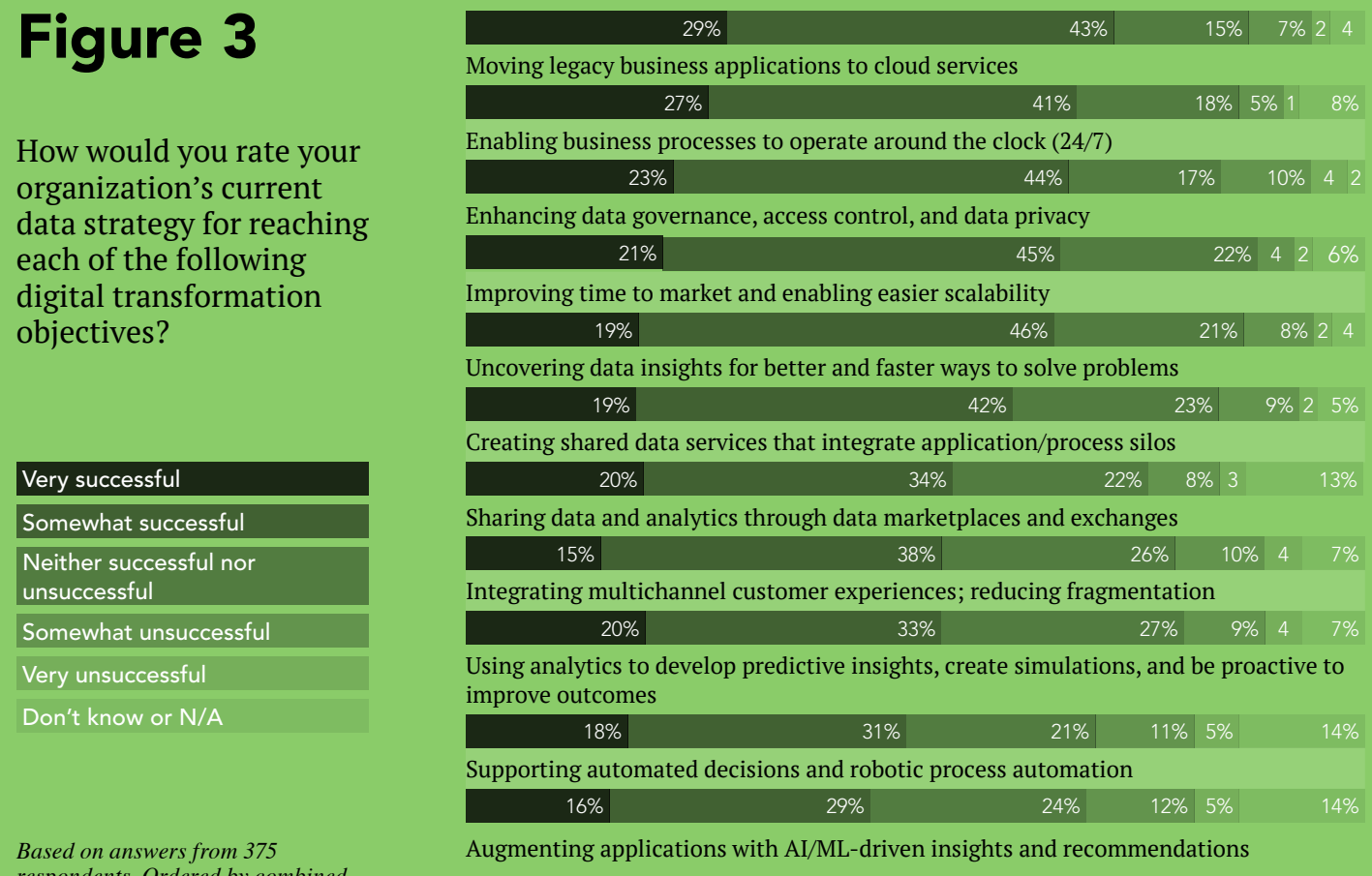
- Articulating data governance policies and priorities, including for improving data quality
- Ensuring alignment with business objectives

We asked participants to rate the success of their organization's current data strategy for reaching several important digital transformation objectives (see Figure 3). The results show that organizations' data strategies are most successful for moving legacy business applications to cloud services; 72% say they are successful.

Of course, some application migrations are more complicated than others. Migrating business-critical applications requires careful planning, including documenting data dependencies between application components. Organizations need to monitor how cloud migration affects data flowing into BI reports, data warehouses, and embedded analytics in applications. Organizations should also address data governance concerns, especially the protection of sensitive data from exposure during migration.

Figure 3

How would you rate your organization's current data strategy for reaching each of the following digital transformation objectives?



Based on answers from 375 respondents. Ordered by combined "very successful" and "somewhat successful" responses.

Enabling business processes to operate around the clock is often an important digital transformation goal; more than two-thirds (68%) say their current data strategy is successful.

Enabling business processes to operate around the clock is often an important digital transformation goal, particularly for global firms that have subsidiaries, development groups, supply chains, manufacturing facilities, and customers in different time zones. More than two-thirds of respondents surveyed (68%) say their current data strategy is successful for handling digital transformation to continuous operations. Modern operational applications are functioning 24/7, and they require a continuously available data infrastructure.

Data strategies also appear to be largely successful for meeting digital transformation goals for improving time to market and enabling easier scalability (66% successful). Modern, agile data platforms have AI-infused automation that helps organizations reduce data latency that slows time to market. Smart automation is also key to modern data integration and data management systems and services.

As part of digital transformation, data catalogs need to support master data management (MDM), product information management (PIM), and customer information management (CIM) solutions—or multidomain MDM that integrates these domains—to ensure that metadata and master data information is coordinated. If an organization is using MDM and/or PIM, data catalog coordination is important to eliminating data confusion that often contributes to delays in product or service rollouts as well as difficulties analyzing business performance.

Modern data catalogs are important to self-service user access to business data. The catalog can provide information about the quality of the data and how it is related to other data. Managing knowledge about the data through data catalogs and higher-level MDM is vital for data architecture scalability as data volumes grow and the data increases in complexity, especially as new sales and marketing channels are added.

AI/ML augmentation is a challenge. Research participants indicate that data strategies are less successful for advanced objectives such as augmenting applications with AI/ML-driven insights and recommendations (17% are unsuccessful and 45% are successful; 24% say they are neither). Augmentation is important to emerging data-driven applications—for example, to facilitate intelligent one-on-one marketing, customer personalization, e-commerce engagement, and fraud detection.

Augmentation with AI/ML also enables BI systems and business applications to deliver in-context prescriptive recommendations to users based on their activity, business rules, and predefined requirements. To enable AI/ML augmentation, data strategies must support technical modernization for scalability and concurrency; strategies must also address data trust and cultural issues such as data literacy and user receptivity to AI-driven recommendations.

Data sharing is an important part of data strategy. Data sharing, which can include services that monetize data and analytics, is important for maximizing the value of data assets. Figure 3 shows that organizations are reasonably confident in their data strategies for improving and increasing data sharing, which digital transformation should enable.

Almost two-thirds of research participants say their data strategy is successful for supporting the creation of shared data services that integrate application/process silos (61% are successful). Over half of participants (54%) indicate that their data strategy is successful for sharing data and analytics through data marketplaces and exchanges.

Self-Service and Data Democratization

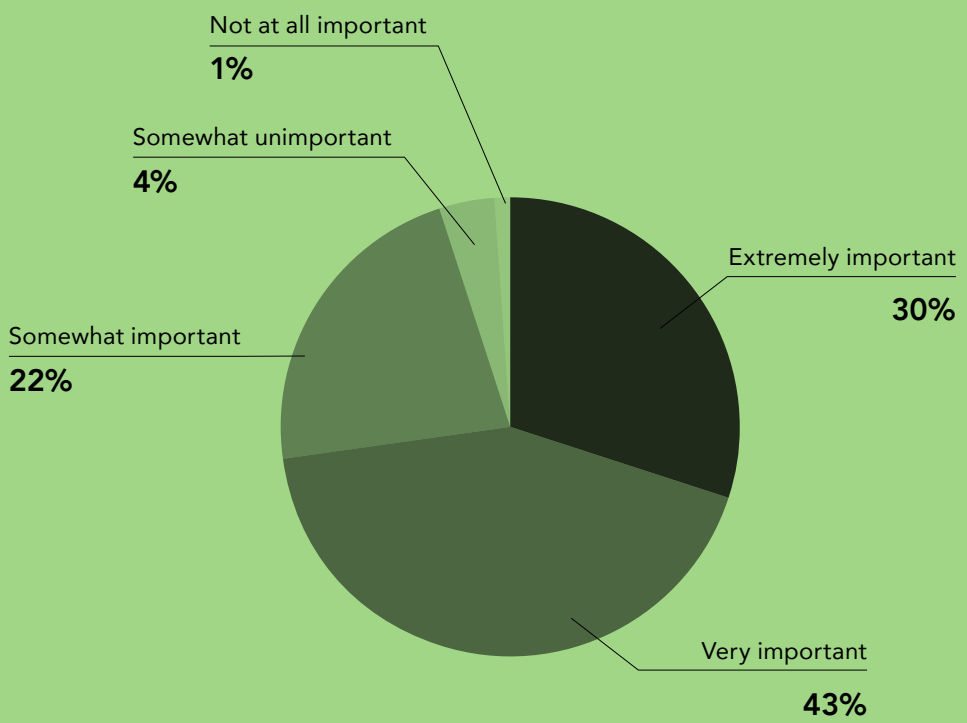
The longstanding trend toward easier, more intuitive user experiences extends beyond personalizing BI dashboards and visual analytics. Many users want self-service functionality for data integration, data pipeline development, interaction with data catalogs, data observability monitoring, and spinning up cloud data management services.

Modern data catalogs enable self-service data consumers to view data quality and data observability information to ensure that data is trusted and safe to use. Data observability, often associated with DataOps, expands beyond data quality monitoring to integrate visibility into other factors and processes affecting the health of an organization's data. The key objective of data observability is to enable faster and easier troubleshooting.

Our research shows that data democratization and increasing self-service functionality are highly important to organizations' data modernization strategies. Only five percent of research participants indicate that it is unimportant, while nearly one-third (30%) call it extremely important and 43% say it is very important (see Figure 4).

Figure 4

How important are data democratization and increasing self-service functionality to your organization's data modernization strategy?



Based on answers from 380 respondents.

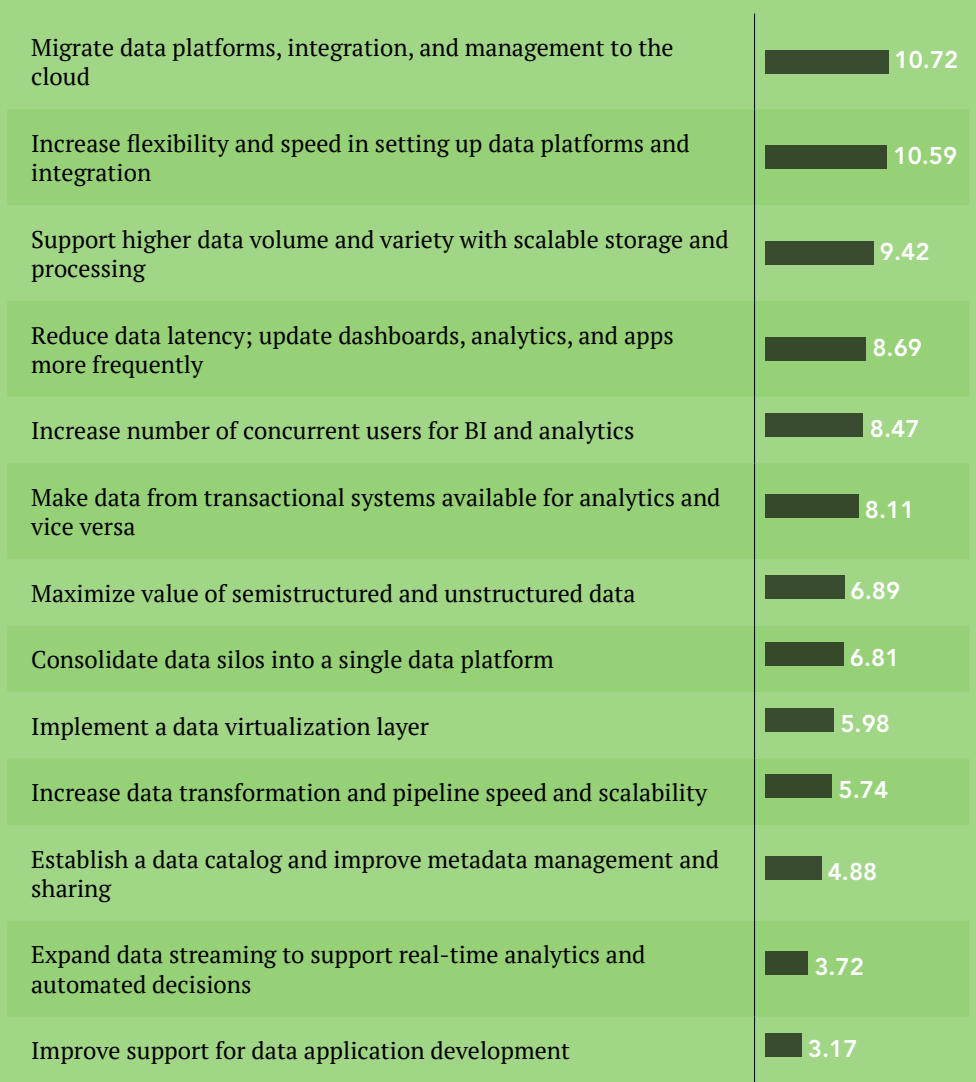
Democratized, self-service users need trusted, relevant data assets to realize optimal value. Organizations need to ensure that self-service data consumption and analytics does not worsen data quality, including data duplication problems that often arise due to increased data silos. Data strategies must balance self-service technology implementation with enterprise data management and governance. As we examine research about data platforms, data integration, and data management, we will discuss how data strategies can better balance self-service with enterprise capabilities and recommend best practices.

Platforms, Integration, and Management Priorities

With our discussion of business drivers accelerating modernization and data strategies as context, we begin a deeper look at experiences, trends, and best practices.

Figure 5

For your organization, how would you rank the importance of the following objectives for modernizing data platforms, integration, and management?



Based on answers from 366 respondents, who were asked to rank the options. Ordered by weighted average.

Cloud migration continues to be a prominent strategic direction. In Figure 5, we can see that migrating to the cloud and gaining one of its key advantages—increased flexibility and speed in setting up data platforms and integration—rank as the two top modernization objectives for organizations surveyed (46% rank cloud migration as their number one objective). Ranked third is another common cloud migration driver: the need to support higher data volumes and variety with scalable storage and processing.

Cloud adoption is often business-driven, which means that chief information officers (CIOs) and chief data officers (CDOs) need to scope adoption (or migration) phases to ensure that projects deliver the desired business outcome. Cloud-based serverless options enable organizations to dynamically provision and scale data platforms in response to business needs. Pay-as-you-go cloud computing models make it easier for organizations to expand or reduce services elastically.

Although newer organizations may have never had anything but cloud computing, most organizations that have adopted a cloud-first strategy for new data systems continue to have a hybrid of on-premises and cloud-based data platforms that have data integration and management capabilities as well as independent data integration and management systems.

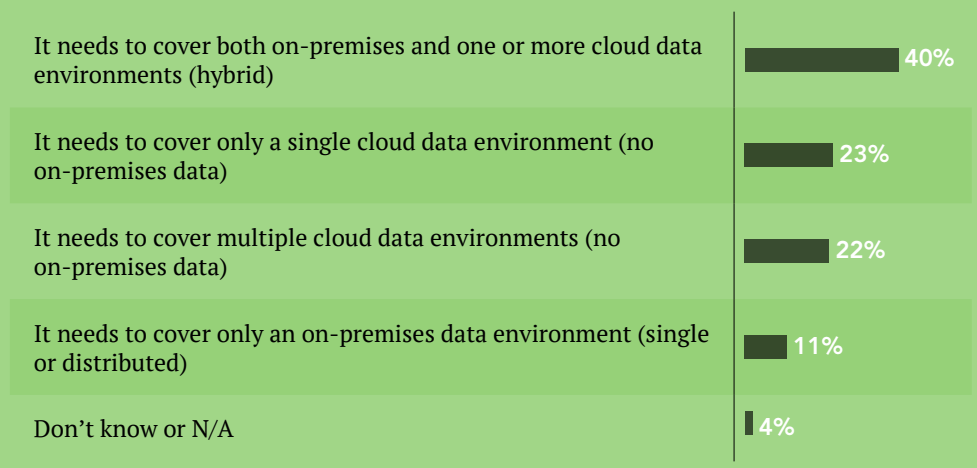
Many organizations adopt multiple cloud platforms for data infrastructure, often to avoid vendor lock-in or a potential single point of failure. As a result, many organizations have hybrid multicloud data environments.

TDWI asked organizations whether their data management and governance need to cover an on-premises data environment only, one or more cloud environments, or a hybrid of on-premises and cloud-based data environments (see Figure 6). The largest percentage indicates that their organizations need to cover a hybrid multicloud data environment (40%). Interestingly, 45% say they have no on-premises data; they have either single or multiple cloud data environments only. Only 11% have to manage and govern only an on-premises data environment.

Figure 6

What does your organization's data management and governance need to cover? (Please select the most appropriate answer.)

Based on answers from 344 respondents.



Hybrid multicloud data environments are challenging to manage and govern. Organizations often must keep data on premises due to data localization laws and other governance requirements.

Hybrid multicloud data environments are challenging to manage and govern. Data distribution makes it difficult to assemble single, integrated views of all relevant data, improve data quality, track data lineage, and monitor data use.

Although some organizations are in the process of consolidating data onto cloud platforms, others must keep data on premises due to data localization and residency laws that require organizations to collect, process, and store data about a nation's citizens or residents in their respective countries. For additional data security and governance reasons, some organizations prefer to keep their most sensitive data on premises.

To address challenges, some organizations are deploying data virtualization layers. Others are using emerging data fabric and data mesh frameworks to establish holistic data management and governance across distributed data environments. This report discusses these options in more detail later.

Overcoming User Satisfaction Issues

Ranking high among modernization objectives shown in Figure 5 is increasing the number of concurrent users for BI and analytics. To enable informed decisions and actions across enterprises, many organizations are migrating from on-premises BI instances that support limited numbers of users to cloud-native services that offer easier user adoption as well as higher concurrency. As more users come online, pressure increases on the data infrastructure to reduce data latency and increase the frequency of updates to dashboards, analytics, and applications. Figure 5 shows that this is a prominent modernization objective.

Satisfied users are those who have the right data at the right time to answer business questions, solve problems rapidly with informed solutions, and innovate through discovery of new data insights. Although users largely interact with data through a BI tool, service, spreadsheet, or business application, it takes the entire data platform stack, including data integration and management capabilities, to satisfy users.

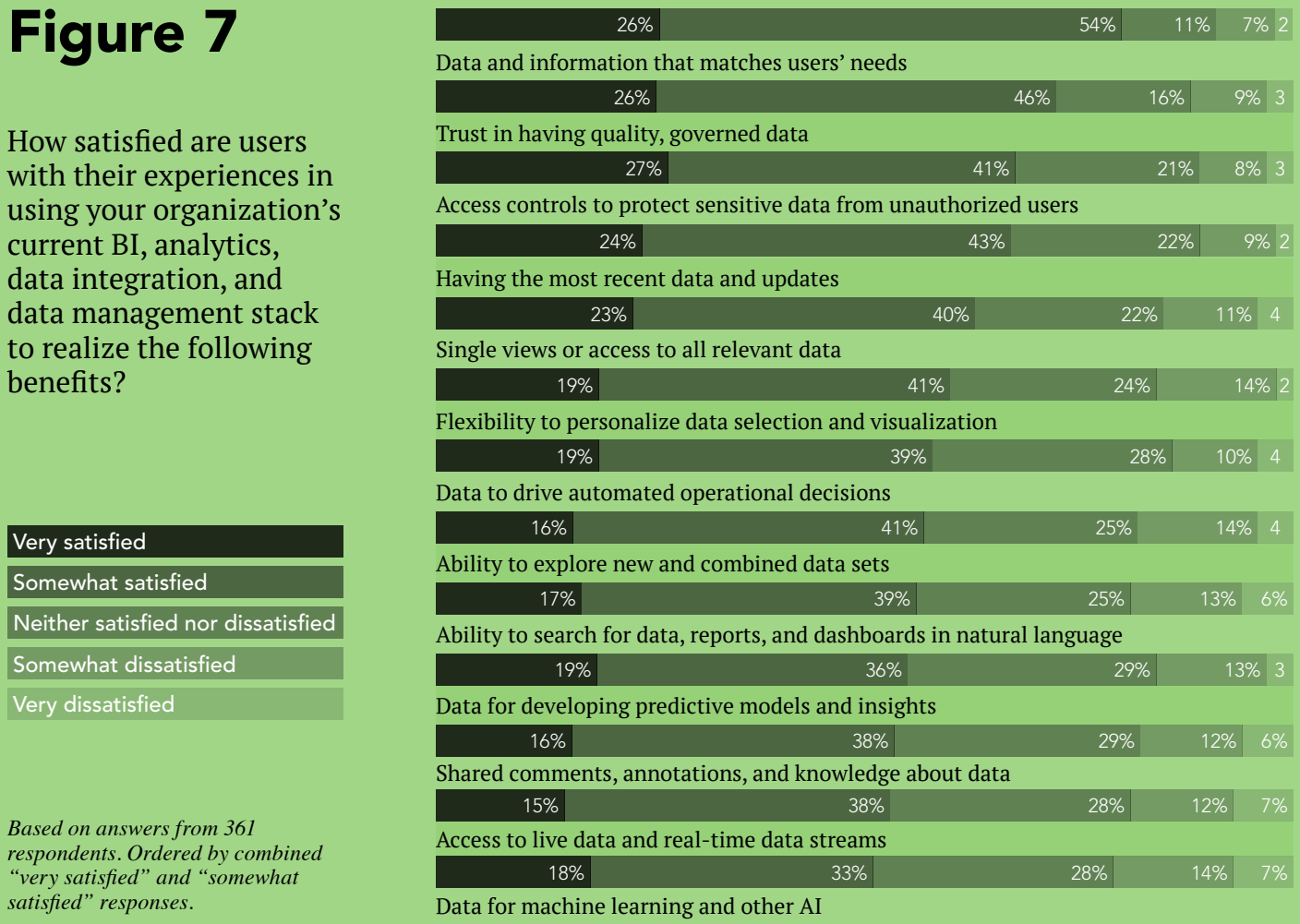
In Figure 7, organizations surveyed report that when using their current BI, analytics, data integration, and data management stack, users are most satisfied with their experience provisioning data and information that matches users' needs (80% say their users are satisfied). This suggests that most organizations in our research are doing a good job of gathering accurate user requirements and ensuring that the underlying semantic layer, ETL, and data warehouse enable access to data fit for users' purposes.

Satisfaction is reasonably strong for gaining single views or access to all relevant data; 63% are satisfied, an improvement over our research one year ago, when 43% said that difficulty gaining a view or access to all relevant data in a single view was a significant hindrance to users making data-informed decisions and realizing value from data assets.[1]

[1] See Figure 2 in the 2021 *TDWI Best Practices Report: Modernizing Data and Information Integration for Business Innovation*, online at tdwi.org/bpreports.

Figure 7

How satisfied are users with their experiences in using your organization's current BI, analytics, data integration, and data management stack to realize the following benefits?



Based on answers from 361 respondents. Ordered by combined "very satisfied" and "somewhat satisfied" responses.

> *Users in most organizations surveyed have confidence in their data; 72% are very or somewhat satisfied with users' ability to trust their data.*

A primary goal of establishing a consolidated data repository such as a cloud data warehouse is to have a single, central source of prepared data relevant to subjects of interest. If consolidation is too difficult or data cannot be moved for governance reasons, heavily distributed data may justify a data virtualization layer. This connects to and queries multiple and distributed sources, aggregates the resulting data, and provides views of it from a single point of access.

Trust in having quality, governed data is relatively strong. Satisfaction falls when users lose confidence in the data and cannot trust their dashboards and analytics. Figure 7 shows that in most organizations surveyed, users have confidence in their data. Nearly three-quarters (72%) are satisfied with how well users can trust the quality and governance of their data.

Data governance rules and policies need to articulate how organizations protect sensitive data, such as customers' personally identifiable information (PII), from improper access and sharing. Users in organizations surveyed appear satisfied with the ability of their access controls to protect sensitive data from unauthorized users (68% are satisfied). Data catalogs are effective for managing where organizations are storing sensitive data and whether users are accessing and sharing it.

Provisioning data for AI/ML and enabling user search are less satisfactory. About one-fifth (21%) of research participants say that users in their organizations are dissatisfied with data for machine learning and other AI techniques. Just over half (51%) are satisfied, which is low compared to the other potential benefits listed in Figure 7. Satisfaction is about the same for getting data for developing predictive models and insights; 16% are dissatisfied and 55% are satisfied.

Data requirements for AI/ML are complex and varied, more so than for BI reporting and dashboards. Setting up data pipelines and data lakes to collect and hold diverse data from multiple sources and then enabling users to find, access, and prepare often terabytes or petabytes of data for their projects typically requires significant expertise and manual work.

Implementing automated tools can increase satisfaction, although sometimes the various automated tools are not well integrated. Rather than use separate data transformation and preparation tools, organizations should evaluate comprehensive (and increasingly cloud-based) data integration and management platforms for analytics and AI/ML projects.

Data catalogs, which may be integrated with these platforms, can play a key role in enabling users to employ metadata to locate relevant data in a cloud data lake or other data sources. Through metadata in the data catalog, users can learn about the data's quality and relevance. AI/ML-driven automation inside enterprise data catalogs supports the scalability to handle large volumes of diverse and changing data and to manage metadata updates automatically.

To gain faster and more complete insights, many business users want interactive data search and discovery using their natural language. They can then examine diverse data sets and improve contextual understanding of what they see in traditional reports and dashboards. However, in Figure 7, 19% of research participants say that users are dissatisfied with their ability to search for data, reports, and dashboards in natural language (56% are satisfied).

The results in Figure 7 suggest that users in many organizations remain limited to traditional and often technically challenging SQL query development to access and interact with data. Some BI and analytics solutions provide natural language processing (NLP)-infused search and discovery capabilities. NLP matches linguistics with processing power to interpret speech and text for meaning about entities (people, places, and things), concepts (words and phrases expressing something), themes, and sentiments. When BI and analytics systems include NLP, users can query data through their natural language. Leading solutions apply machine learning to increase the speed, flexibility, and precision of answers.

When integrated with modern semantic layers and data catalogs, users' search and discovery can become more accurate and comprehensive. Some integrated solutions augment traditional

querying and visualization with AI/ML-driven recommendation engines to shorten the time it takes users to locate the right reports or data sets.

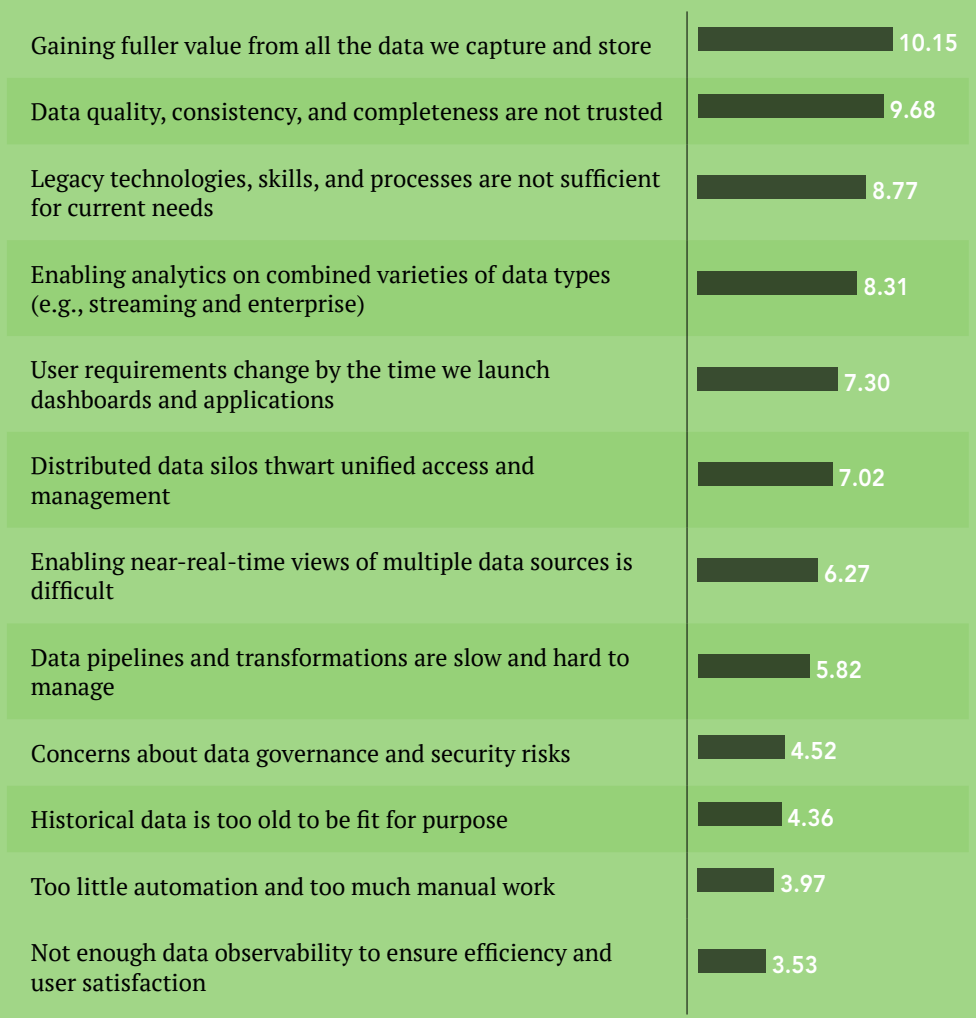
Using knowledge about the data collected in data catalogs and accessible via semantic layers increases users' flexibility in personalizing data selection and visualization. In Figure 7, 60% of organizations surveyed say their users are satisfied with personalization of data selection and visualization.

Top Challenges in Maximizing the Value of Data

User satisfaction is an important ingredient in maximizing the value of data assets and is thus a key measure of the influence of the organization's data platforms infused with data integration and data management capabilities as well as independent data integration and management tools and services. Today, it is not just users who need to receive the right data at the right time. Automated, data-driven algorithms and applications also require continuous, curated, and timely data flows.

Figure 8

What are your organization's biggest challenges in enabling users, algorithms, and/or applications to maximize the value of data assets?



Based on answers from 349 respondents, who were asked to rank the options. Ordered by weighted average.

Gaining more value from all the captured and stored data is the biggest overall challenge, according to our research.

We asked research participants about their biggest challenges enabling users, algorithms, and/or applications to maximize the value of data assets (see Figure 8). Simply "gaining fuller value from all the data we capture and store" was the biggest challenge by consensus; 46% selected this as their top challenge).

Organizations today capture and store a variety of data types, including raw, unstructured, streaming data and structured enterprise data from multiple sources. Figure 8 shows that enabling analytics on varied data is a major challenge; it ranks fourth. Organizations need to overcome this challenge to advance with both data science and modern, analytics-infused applications.

Organizations surveyed indicate their frustration with the insufficiency of legacy technologies, skills, and processes for current needs; they rank this challenge third. Legacy systems' inflexibility is a modernization driver; we see in Figure 8 that the problem of user requirements changing by the time dashboards and applications are launched is a significant challenge. Thanks to cloud computing, data management can be less fixed in size and scale, which means that organizations are able to respond to changing user requirements with cost-effective agility and meet dynamic business analytics needs.

Data trust, quality, consistency, and completeness are issues. Looking at specific data challenges, the lack of trust in data quality, consistency, and completeness ranks second; 20% of respondents rank it first and 30% rank it second). Distributed data silos exacerbate problems with data quality, consistency, and completeness because each silo may have different data and define and organize data differently. Distributed data silos can thwart unified data access and management; this challenge ranks sixth in Figure 8.

Semantic layers have long been critical to the quality and consistency of business representations of data for BI reporting, calculations, and OLAP multidimensional modeling. Some organizations are looking to solve distributed data quality and consistency difficulties by enhancing the capabilities of semantic layers through use of AI/ML, knowledge graphs for mapping data relationships, and data catalogs.

Organizations are also setting up data virtualization layers integrated with data catalogs to make it faster and easier to view and govern disparate data and uncover data inconsistencies. Finally, where possible, other organizations are consolidating data from silos into unified, cloud-based data platforms so users, algorithms, and applications have a trusted, centralized resource.

Data pipeline developers should consider incorporating data quality steps to save users time fixing the problems later, which is usually done in an ad hoc fashion.

Organizations should enhance data pipelines with data quality. When data quality is a concern, data pipeline developers should consider incorporating data quality steps to save analysts, developers, and data scientists from having to fix problems later. At that point, remediation is usually done in a less systematic, more ad hoc fashion. Data catalogs enable users and administrators to discover data quality problems before pipelines move or replicate data downstream to target destinations.

Data lineage tracking in data catalogs and related processes for data quality, profiling, and validation require techniques such as AI/ML and automation to keep pace with data volume and diversity.

Problems with data pipelines and transformations being slow and hard to manage ranked in the middle of the challenges listed. Unfortunately, legacy code for data acquisition and other data integration jobs forms a large part of the technology "debt" that CIOs and CDOs have to continue to manage as they try to turn their focus (and budgets) toward modernization.

With data proliferating faster and more voluminously across dozens, if not hundreds, of sources, organizations struggle with the central function of data pipelines: data acquisition from multiple sources. *Data acquisition* is the process of identifying and documenting disparate data sources, then ingesting raw data from these sources and transforming it into a usable state for analytics, dashboards, reporting, and other needs.

Bottlenecks and choke points can form in data pipelines and ETL jobs, which may be so numerous that it is hard to determine the source of delays and remove them. Allowing data pipeline processes to access a data catalog or other metadata management in an automated fashion can enable users, algorithms, and applications to find data faster—whether the data is located in a physically centralized data lake or in a table managed in one of multiple distributed data platforms.

Modifying or adding new data remains a long process. Fresher data and quicker updates contribute to timelier decisions and more efficient business operations. With growth in analytics, more users and data applications need access or views of new or updated data as soon as possible.

Our research shows that the largest percentage of organizations surveyed (32% of 338 respondents; not shown) take between one and three months to modify existing data or add new data to data platforms to make it available for users' reporting, dashboards, and analytics. Almost half (48%) can refresh existing or add new data in one month or less, but 16% take longer than three months. However, larger organizations surveyed—those with 10,000 or more employees—are faster; 60% refresh existing or add new data in one month or less with 37% reporting that it takes just two weeks or less. These organizations likely have more technology resources and skilled personnel available to speed data modification and addition.

Users need data refreshes so historical values in the data warehouse or analytics platform are in accord with source data values. Thus, it is important to focus on eliminating bottlenecks, delays, and disconnects in data integration processes and, where appropriate, move toward real-time data access and views. Automated data integration tools are helpful. However, there is no single answer; different user requirements demand different solutions. Organizations need multiple options for reducing latency. We will discuss options later in the data integration section of this report.

Strategies for Modern Data Applications

Operational business applications today need to be data-rich, if not data-driven. As noted, the traditional divide between BI and data warehousing on one side and operational applications on the other does not serve user requirements that depend on deeper integration between data, analytics, and application processes. Unlike rigid business applications of the past, modern data applications need flexibility and scalability to manage and derive value from big data. This highlights the advantages of flexible NoSQL data management, cloud computing, and componentization.

> *For modern data applications, many developers have taken advantage of open source database innovations to overcome the constraints of traditional systems.*

To create data applications, many developers have taken advantage of open source database innovations including document, key-value, column, and other NoSQL systems. These technologies let developers overcome the hardware, storage, and schema constraints that have characterized traditional applications built on relational database management systems (RDBMSs). Developers have been able to optimize their chosen data management system to fit application demands. With the latest versions of these systems running in the cloud and using containerization and open source Kubernetes for orchestration, developers have the flexibility to change database components as needed to meet business requirements.

Embedded analytics and visualizations such as dashboards developed in-house or by third parties are often the essence of what makes one application better than another. Today, developers can embed cloud-native services that bring highly visual predictive insights, metrics, real-time notifications, and data-driven recommendations into larger applications to improve both human and automated decision-making.

Employees in customer-facing operations such as sales and support can take advantage of timely and relevant data insights inside purpose-built applications to improve sales strategies, marketing campaigns, and customer engagement. Many organizations embed analytics into their websites to sense patterns and learn from online behavior. They can use analytics to drive automated online customer engagement, including by determining the appropriate content to present consumers to guide real-time cross-sell and up-sell offers.

Organizations are also using modern data applications to provide data- and analytics-rich portals for business partners (such as in a supply chain, a collaborative manufacturing group, or a healthcare provider network) to share insights for improving operations or dealing with unanticipated problems. Using systems such as data catalogs to make it easier to apply relevant data governance and regulatory constraints, data sharing ecosystems such as these enable organizations to realize additional value from their data and move faster to tap external data sources. Some data sharing ecosystems involve data brokerages that provide aggregated data through portals or embedded BI while others offer reciprocal data sharing among partners.

Seamless access to diverse data sets and sources is most critical. TDWI asked research participants how their organizations would rank the importance

of a series of objectives for delivering rich data insights embedded inside applications and/or to drive autonomous actions. The top objective among most participants is to provide seamless access to diverse data sets and sources (52% of 340 respondents rank this first, not shown).

Data applications are by definition data-hungry; the more organizations can eliminate bottlenecks and delays preventing broad data access, the better. Applications that feature embedded analytics or use AI/ML to inform and guide automated decisions need data platforms that offer continuous and reliable data access for potentially hundreds of thousands of customers around the globe, not just a traditional selection of internal employees. Particularly for nontechnical users, data applications that provide intuitive search, visualizations, and analytics data interaction deliver advantages over legacy applications that require writing queries. The objective of embedding intuitive search, visualizations, and analytics data interaction ranks fifth out of 10 possible objectives.

> *Creating competitive differentiation with richer application experiences ranks second among data application objectives for most research participants.*

Developers seek differentiation with richer, personalized application experiences. The second- and third-highest objectives among research participants indicate the importance of data insights in delivering uncommon user experiences that increase satisfaction. Creating competitive differentiation with richer application experiences ranks second and delivering timely, personalized user experiences, including recommendations, ranks third.

To achieve these objectives, developers need data platforms that support more than just static reports and dashboards that merely await users' actions. In "active" data applications, AI/ML-augmented features drive prescriptive recommendations for next steps and automatically anticipate user requirements.

Giving customers smart, real-time data insights is important. Most CIOs and CDOs tell us that the speed of digital business is faster. Their organizations cannot afford IT being a limiting factor in their ability to operate and make decisions at the speed of digital business. Getting IT out of the way is one of the main drivers behind demands for self-service data access.

In most organizations, legacy applications developed and managed by IT typically provide historical reports that are of limited value. Business users, including partners, increasingly need current information, such as when an order will arrive or whether the status of machinery in a manufacturing process indicates the need for immediate maintenance. Giving data application customers real-time insights ranks fourth on the list of objectives.

With modern data platform support, developers can embed applications with smarter alert capabilities. Alerts can be limited to those relevant to users' concerns and interests instead of bombarding users with notifications, which might cause users to ignore the important alerts.

Addressing Data Application Development Challenges

Most organizations surveyed by TDWI (79%) are satisfied with their ability to develop homegrown applications or deploy third-party applications that use data effectively, deliver actionable insights, and/or enable autonomous actions (see Figure 9).

To gain a more in-depth understanding, we asked how challenging it is for developers to build and deploy applications that work with organizations' current data platforms, integration, and management to provide a variety of important capabilities (see Figure 10). The most challenging task is to enable low-latency querying, search, and analytics. About three-quarters of respondents (72%) say it is challenging. These research insights align with the importance of the objective mentioned above for giving customers real-time data insights. Clearly, developers regard shortening the path to valuable search and query results and delivering timely analytics as critical—but challenging—objectives.

Figure 10 shows that organizations face difficulties using their current data platforms, integration, and management to handle real-time data streaming. More than two-thirds of research participants (68%) say their developers face challenges processing streaming data and change data capture updates.

Nearly as many (64%) struggle to integrate both streaming ingestion and fast, high-volume queries as well as to combine historical and real-time data for complex analytics (also 64%). Streaming data is often set up to arrive continuously from thousands of sources such as IoT sensors, log files, financial trading systems, social networks, and in-game player activity. Organizations may land streaming data in a data lake or perform analytics on streaming data in motion. To perform time-series analysis and uncover important patterns and anomalies, organizations frequently need to integrate the data and analyze it alongside historical data.

When data is landing in different systems and bridging the data silos is difficult, the problem

Figure 9

How satisfied is your organization with its ability to develop homegrown applications or deploy third-party applications that use data effectively, deliver actionable insights, and/or enable autonomous actions?

Based on answers from 343 respondents.

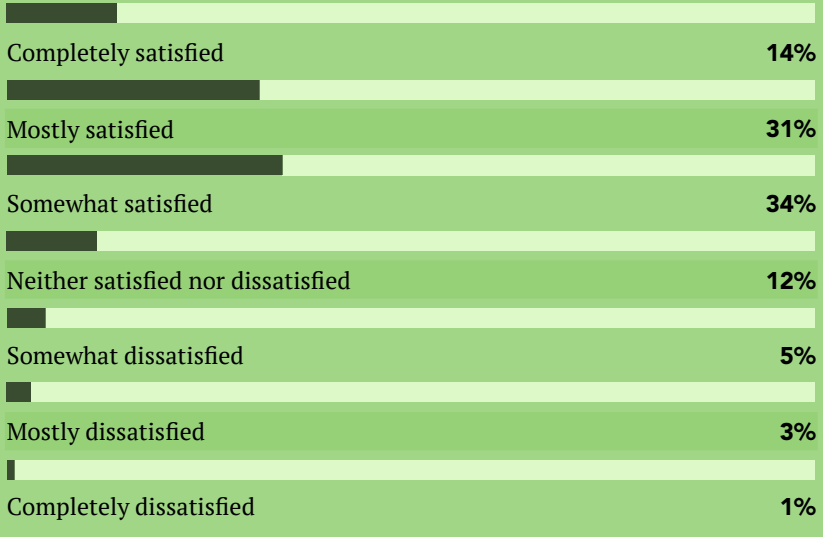
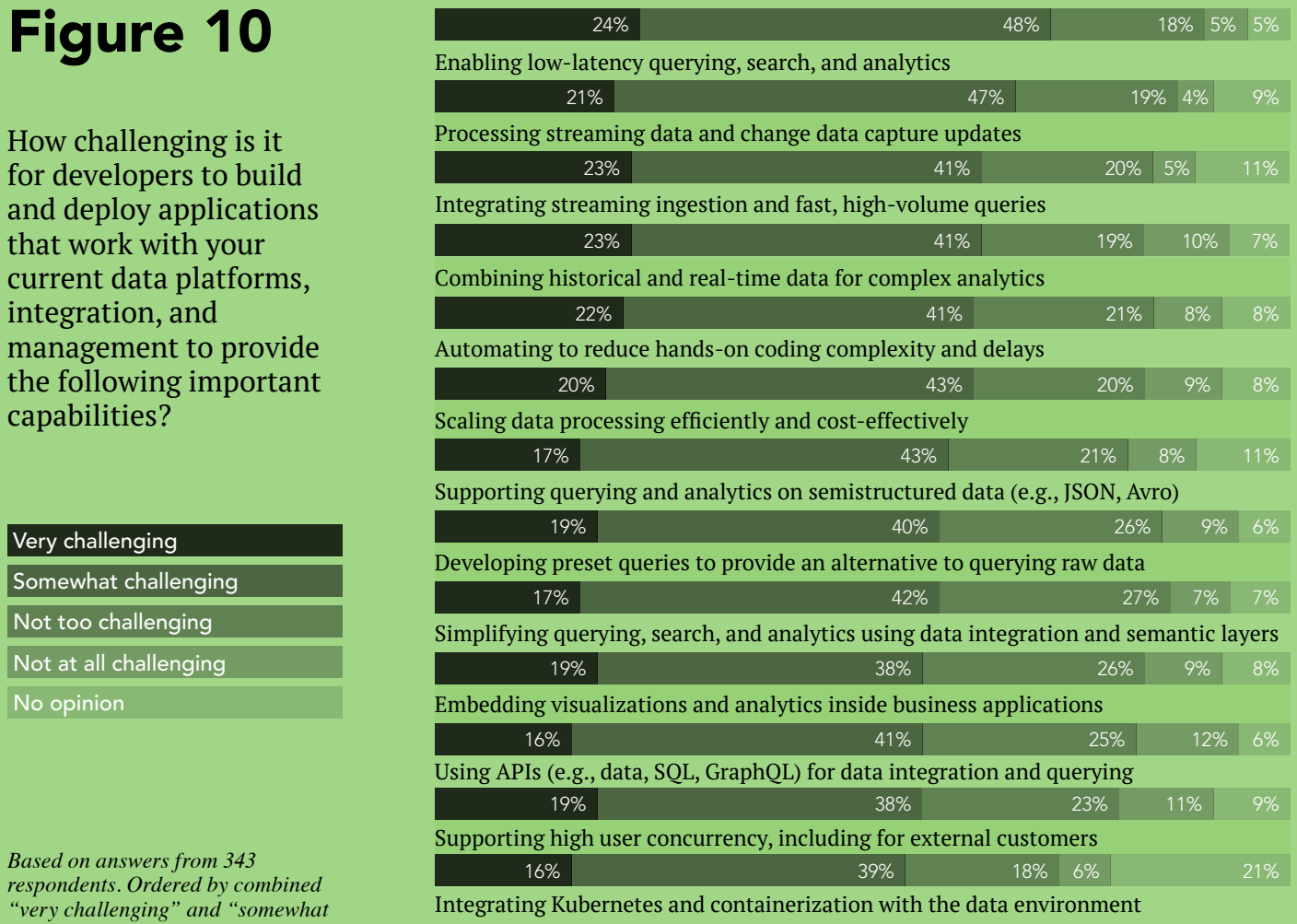


Figure 10

How challenging is it for developers to build and deploy applications that work with your current data platforms, integration, and management to provide the following important capabilities?



Based on answers from 343 respondents. Ordered by combined "very challenging" and "somewhat challenging" responses.

introduces latency. Users typically need the most current data and want to enrich analytics with new variables. A best practice is to develop a data strategy that is flexible for different business use cases. For some cases, virtualized views of integrated data sets in a virtual data warehouse is the fastest path to complete views of the data; in others, the best course is to consolidate data into a single, unified, highly scalable cloud data platform.

Organizations face difficulties reducing coding and simplifying search and access. Figure 10 shows that developers in many organizations are looking for low- or no-code solutions that alleviate the need to perform custom data access and integration coding. Nearly two-thirds (63%) say that automating to reduce hands-on coding complexity and delays is challenging. Nearly as many research participants (59%) say it is challenging to simplify querying, search, and analytics using data integration and semantic layers.

Modern data integration and semantic layers are key to supporting data applications as the data grows in size and complexity. Automated mapping and integration with metadata intelligence

contained in data catalogs can shield users from underlying complexity and support trends toward using low- or no-code development for building, testing, and deploying applications.

> *Developers are looking for low- or no-code solutions that alleviate the need to perform custom data access and integration coding.*

Organizations should evaluate solutions that enable the creation of predefined data integration routines and preset queries, which provide alternatives to querying raw data. Our research indicates this is challenging for more than half of organizations surveyed (59%). Modern data integration and semantic layer solutions use automation along with predefined and preset functions to overcome challenges.

Scaling data processing efficiently and cost-effectively is challenging. In Figure 10, nearly two-thirds of research participants (63%) indicate that processing scalability is an issue for data application development. As data applications grow to depend on access to bigger and more varied data volumes, organizations need data integration and data platforms that can scale on demand rather than be limited to fixed data volumes and processing power. Cloud data platforms enable organizations to address limitations, including through separation of storage and computational processing that allows organizations to scale each independently, adding flexibility.

Data Management for Different Use Cases

Data management systems are essential to maximizing the value of data. Data platforms implement different types of data and document management; they have to balance stability with flexibility as data environments evolve, users' data interaction modes change, and developers create new types of data- and AI/ML-driven applications. Technology innovation is ongoing as organizations seek to push past legacy system constraints and capture value from new types of data and huge data volumes. Data management systems must increasingly satisfy demands for faster data updates and real-time data.

> *Data platforms must balance stability and flexibility as data environments evolve, users' data interaction modes change, and developers create new types of applications.*

TDWI asked research participants about the types of data management systems, platforms, storage, or services they currently use and plan to use in the future, either on premises or in the cloud. We also asked about what types of use cases their organizations support with their selected data systems. We asked the same questions regarding prevalent types of data integration, which we will discuss later in this report. We asked about these use cases:

- **Data science and AI/ML.** Data science projects often involve exploratory, predictive, and real-time analytics. Data scientists develop

and test AI/ML models and algorithms that depend on access to hundreds of terabytes (if not petabytes) of data contained in a data lake, data warehouse, or unified platform that combines them. Data scientists often want to access unstructured and semistructured sources such as log files, social media, customer and consumer experience data sources, and mobile device data along with structured sources. Data scientists and analytics application developers use languages such as Python, R, and Scala and ML libraries in addition to SQL.

- **BI, OLAP, dashboards, and business analytics.** Workloads for these use cases typically rest on well-established data transformation (ETL and ELT) and data warehouse processes. Most users primarily need structured and transformed data, SQL query-and-reporting functionality, some search and NLP, and self-service visualization (such as a dashboard). BI/OLAP systems support dimensional modeling and online analytics processing (OLAP) for consolidation, drill-down, and slicing and dicing data interaction to gain different and deeper perspectives on the data. More recently, BI/OLAP users can access data lakes through semantic layers integrated with data catalogs.

- **Applications (e.g., operational, e-commerce).** These use cases typically demand high availability and concurrency, frequent updates (if not real-time data), and high data quality. The previous section of this report detailed the emergence of data applications, which rely on AI/ML to drive outcomes such as automated decisions, personalized recommendations, content delivery, and predefined responses to trends and situations.

- **Customer or partner engagement or data sharing.** *Data sharing* means giving internal data consumers in different departments—as well as external partners, customers, and other interested parties—self-service access to centrally managed data sets. Data marketplaces and exchanges are gaining traction as a means of sharing data, reports, dashboards, and analytics built on top of data warehouses and data lakes. Many organizations monetize data by providing services in a data marketplace or exchange. Thus, most organizations carefully control data sharing workloads. These vary from data consumption through portals and dashboards to deeper analytics.

As we discuss the research survey results about current and planned usage (see Figure 11), we will also describe associated survey results about the above use cases. Note that because data management systems or platforms often support multiple use cases, respondents could indicate more than one use case for each.

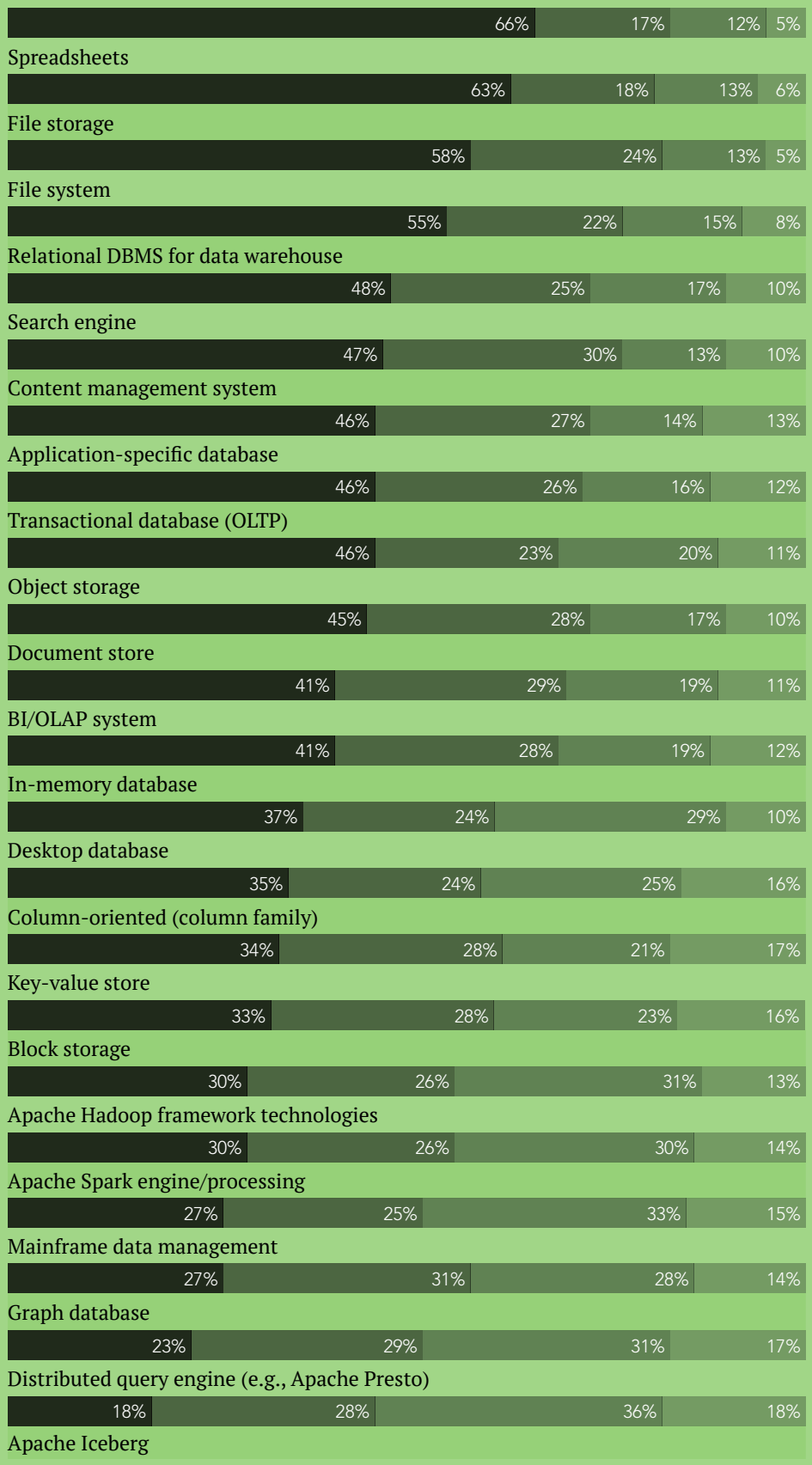
Not surprisingly, spreadsheets show the highest percentage of current use (66%). Spreadsheets remain a common tool for individual users to extract data from sources and organize, categorize, manipulate, and analyze it. However, only 17% plan to use them as a data management system, platform, storage, or service in the future, which suggests that many respondents are interested in switching to more robust and purpose-built technologies and services. Basic file storage and file systems show similar results in the research.

Over half of respondents (55%) are using a relational DBMS and 22% plan to use one; these results are similar to those from last year's TDWI Best Practices Report research survey.[2]

[2] Ibid., Figure 4.

Figure 11

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



Based on answers from 336 respondents. Ordered by "currently using" responses.

As a well-established, fundamental technology and data model, our research shows that RDBMSs support a spectrum of use cases. The highest percentage use an RDBMS for BI, OLAP, dashboards, and business analytics (44%, figure not shown; research participants could check all use cases that applied). More than one-third use an RDBMS for applications (38%). One-quarter (25%) use one for data science and AI/ML and 21% for customer or partner engagement or data sharing.

An RDBMS is often used, but not exclusively, as the DBMS for online transactional processing (OLTP). OLTP systems manage execution and recording of database transactions for a vast range of applications, including banking, e-commerce, and travel booking. Nearly half of organizations surveyed (46%) are using an OLTP system currently and 26% plan to use one.

Today, OLTP systems must scale to even higher levels of transaction intensity with continuous availability and fault tolerance. Some organizations integrate analytics and AI/ML with OLTP to gain real-time insights for fraud detection, for example, and to drive smarter, more personalized e-commerce. In our research, the highest percentage are using OLTP systems for applications (43%; not shown); these likely include enterprise resource management. More than one-third (35%) are using OLTP systems for BI, OLAP, dashboards, and business analytics; users in this scenario likely need direct access to live business data generated by the application. Just under one-quarter (22%) are using OLTP for data science and AI/ML, and 18% are using it for customer or partner engagement or data sharing.

NoSQL for Modern Data and Application Demands

RDBMS data modeling constraints and scalability limitations for big data have made room in the market for growth in NoSQL data management such as column-oriented databases, document databases, graph databases, and key-value stores. Open source has been the locus of NoSQL innovation. Most types of NoSQL databases today run on premises and in the cloud or across both types of platforms. Here, we explore what the research shows for prominent NoSQL systems.

Column-oriented databases. A column-oriented database stores data tables by column rather than by rows the way a traditional RDBMS stores data. Figure 11 shows that 35% of organizations surveyed currently use a column-oriented (or "columnar") database and 24% plan to use one. Column-oriented databases can deliver faster performance for more complex analytics than standard RDBMSs; column storage reduces the amount of data retrieved to satisfy a query.

> *Column-oriented databases can deliver faster performance than standard RDBMSs.*

Most column-oriented DBMSs use indexes and compression to further reduce disk I/O and the amount of data read in response to a query. Column-oriented databases on cloud platforms take advantage of massively parallel processing and in-memory computing to provide scalable high performance. One-fifth of organizations surveyed (20%) are using a column-oriented database for data science and AI/ML; somewhat more are using one for BI, OLAP, dashboards, and business analytics (33%; not shown). One-third of research participants (33%) use a column-oriented DBMS

for applications and 15% use one for customer or partner engagement or data sharing.

Document databases (or stores). Document databases store values in documents, which traditionally have included a tremendous variety of items such as blogs, social media, online customer comments, and other content. Document databases also tend to address a multitude of use cases with expressive query languages and support for secondary indexing. This allows document databases to run a wide variety of queries to support both transactional workloads and the aggregations/computations that enable more real-time, automated analytics found in modern applications.

Key-value databases (or stores). Using a simple data model of keys and values, key-value databases let users choose keys and query values stored in the database. The simplicity and flexibility of these systems make them attractive to developers of data applications. Figure 11 shows that a significant percentage of organizations surveyed (45%) are currently using a document database (or store) and 34% are using a key-value database (or store); 28% plan to use each of these types of databases in the future.

> *The largest percentage of organizations surveyed (37%) are using document stores for applications.*

Regarding use cases, the largest percentage are using a document database for applications (37%; not shown). This is an indication of their popularity with developers. Nearly a third (30%) are using a key-value store for applications. About one-quarter (26%) are using a document database for BI, OLAP, dashboards, and business analytics and the same percentage are using a key-value database.

Graph databases. Graph databases are effective for enabling easier discovery and analysis of data relationships. Graph databases and query languages manage and retrieve data relationships. A graph database offers an alternative to an RDBMS for storing complex data and exploring and analyzing data relationships across large numbers of data points. Just over one-fourth of organizations surveyed (27%) are currently using a graph database and 31% are planning to use one. These percentages show an increase over 2021 percentages (19% and 23%, respectively).[3]

The largest percentage are using a graph database for BI, OLAP, dashboards, and business analytics (29%), although nearly as many (26%) are using one for applications (figure not shown). With a graph database, developers can use special query functions to bypass writing complex SQL statements. This enables users to discover associations in relational databases. Developers can program custom routines to convert graph structures into relational structures.

Data Management Modernization Objectives

Figure 12 shows that improving data quality is the top-ranked objective for most organizations in modernizing their data stack. The stack could be made up of a data platform that incorporates data integration and management capabilities, or it could be connected to independent data integration and management systems or services.

Nearly two-thirds (60%) say data quality is their top objective. Data quality goes together with data governance, which most participants rank second (45% rank it second and 13% rank it first).

[3] Ibid.

Data governance should cover both defensive and offensive concerns. Dominating defensive concerns is regulatory adherence; organizations need to implement rules and policies for protecting sensitive data such as customers' personally identifiable information (PII). Offensive concerns are about improving the value of data assets through data curation. Data curation processes include those for improving data quality, accuracy, consistency, and completeness.

Not all use cases demand the same degrees of defensive or offensive data governance. For example, dashboards and reports for monitoring changes in the data over time need offensive data governance to curate accurate and consistent data so trends and comparisons drawn are valid. Defensive data governance constraints should be in place to control access to sensitive data via dashboards and reports.

Requirements of analytics and AI/ML often depend on the project; data governance needs to be flexible. Some projects may need raw data streams with no data curation and others may need only specific curation processes such as data profiling

Figure 12

How would you rank the importance of the following data management objectives for driving your organization's modernization of its data stack?

Based on answers from 336 respondents, who were asked to rank the options. Ordered by weighted average.

and targeted data quality corrections and cleansing. Defensive data governance constraints, preferably automated in data pipelines, should guide developers and data scientists so they do not view or share sensitive data inappropriately.

Increasing the automation of data management ranks third. Data governance and curation are traditionally heavily manual processes, as are many other data management steps. This can result in data management errors, inconsistencies, delays, and higher costs. Our research shows that organizations are strongly interested in increasing data management automation; it is the third-highest-ranked objective.

Automation is important for scalability as data volumes and diversity increase. Organizations need data management, data integration, and data cataloging automation, especially tools infused with AI/ML. This enables them to systematize and scale version control, reuse, and workload change management, which Figure 12 shows is important to some respondents.

> *Research results indicate that organizations are interested in looking beyond traditional databases to maximize the value of all data assets.*

Organizations want to gain advantages of systems beyond traditional RDBMSs. The fourth and fifth-ranked objectives show significant interest in looking beyond traditional systems to maximize the value of all data assets. This includes taking advantage of in-memory processing for faster queries and scalable, massively parallel processing (MPP).

As data volume and velocity rise, organizations' data ingestion processes are under pressure. Along with familiar big data files such as IoT sensor data and log files, application programming interfaces (APIs) are making new, typically JSON- or Apache Avro-typed data sources easily available (note that using APIs to integrate data flows and tools ranks seventh). Through data pipelines, organizations need to move big, often semi- or unstructured data from sources to a target, typically a cloud data lake, for storage and easier access for analytics.

In Figure 12, supporting ingestion of large data streams ranks fourth among objectives. This suggests that organizations are evaluating tools that offer more automated and systematic scheduling and monitoring to ensure complete ingestion and that can support higher velocity pipelines.

Ranked fifth is using NoSQL to deliver data insights faster than with a traditional RDBMS. NoSQL databases give data scientists, developers, and business users alternative options; they are not limited to moving data through traditional structuring and transformation processes. These processes may be either unnecessary or too difficult and slow for their use cases. NoSQL adds flexibility to how users can query the data, including the use of languages and extensions that may be a better fit for workloads than standard SQL.

Our research shows that many organizations need to focus attention on modernizing connectors and improving use of data catalogs and semantic layer technologies to ensure quality, standardization, and to promote reuse and less duplication. For example, in Figure 12, we see that improving data indexing for faster query performance is ranked eighth among objectives. Indexing is an important performance optimization technique for locating data faster and more efficiently and reducing data retrieval. For some organizations, it may merit a higher position among objectives.

Data virtualization is useful for trusted data views drawn from multiple sources. Ranked sixth among objectives is using data virtualization to access and govern multiple sources. Particularly as hybrid, multicloud data environments grow more common, organizations need to access and govern multiple data locations.

A data virtualization layer can improve how organizations govern data access by serving as a single point of entry to data sources. Organizations can integrate single sign-on, authentication, and encryption with the data virtualization layer so only authorized users can access and consume sensitive data that the organization needs to secure. Users still have the flexibility to use their preferred BI and analytics tools, but they can only access the governed data through the virtualization layer.

Cloud Computing Practices for Accelerating Value

In this report thus far, we have seen that a high priority for most organizations is to shift the center of data management and integration to the cloud. This entails significant cloud data migration from legacy on-premises data platforms to cloud data lakes and data warehouses, often onto multiple cloud providers' platforms.

Phased migrations are common, in which organizations migrate systems incrementally to the cloud over time. This approach has the benefit of limiting disruptions so unrelated workflows continue and data remains available for other users across the enterprise. However, by leaving some systems on premises and others in the cloud, data access and integration challenges become more complex.

Phased cloud migrations are common; this approach has the benefit of limiting disruptions so unrelated workflows continue. However, data access and integration can become more complex.

Hybrid multicloud environments are often established by design. Organizations have to adhere to data localization and residency laws that require organizations to collect, process, and store data about a nation's citizens or residents in their respective countries. Multicloud is also common for data management because organizations have multiple cloud platform options and may choose to take advantage of different providers' strengths and want to avoid vendor lock-in.

Integrating views of hybrid multicloud data is a top concern. In answer to our question about which issues are most critical to accelerating value from the use of the cloud for data integration and management, the majority of organizations rank connecting and integrating views of on-premises and cloud-based data first (see Figure 13; 51% rank it first). Access and sharing of data located in multiple cloud platforms ranks second.

These results and others in Figure 13 highlight the need for holistic data management, integration, and governance to realize the full potential of cloud computing for empowering users with complete data views and access regardless of whether the data is located on premises or in the cloud. Some organizations may choose to accelerate data consolidation in the cloud to avoid having too many data silos. Figure 13 shows that increased data silos due to unplanned cloud adoption ranks fourth among concerns.

Organizations want to enable broad data access and run data integration and transformation routines across the hybrid multicloud data environment without increased manual coding. Automated data acquisition, integration, and transformation tools play a key role in avoiding the delays, costs, and complexity of manual coding. Along with automation, organizations should evaluate tools that offer templates for predefining common data access and integration jobs. This helps organizations scale up workloads and support rapid growth in concurrent users, an issue that ranks third in Figure 13.

To reduce data movement and the need for specialized data integration programming, some organizations are adopting data virtualization, data fabrics, and data mesh strategies to enable data views and governance across distributed environments. These strategies could help organizations stay on top of data ingestion, egress, extraction, and migration costs, which

Figure 13

How would you rank the importance of addressing the following issues critical to accelerating the value of your organization's use of the cloud for data integration and management?

Based on answers from 325 respondents, who were asked to rank the options. Ordered by weighted average.

ranks sixth among issues critical to accelerating value in Figure 13.

Organizations need better cost visibility and observability. Reducing costs is often a major goal of cloud data migration, but many organizations struggle to control and manage costs. Figure 13 shows that lack of visibility and metrics for managing total costs is a challenge (ranked seventh). Migration to multiple cloud platforms adds more options and cost combinations to track. To address these challenges, organizations should evaluate tools that provide continuous, automated monitoring as well as intuitive interfaces so both business and IT managers can understand cost drivers.

> *To increase visibility into data problems and their business impact, some organizations are employing data observability practices and tools, often as part of DataOps.*

Some organizations are employing data observability practices and tools, as noted earlier, often as part of DataOps. Data observability extends the idea of complete data governance to enabling organizations to understand the health of data environments and be able to address issues quickly. Key areas of focus include issues causing downtime, poor data quality, and data acquisition and integration bottlenecks. As with data governance, a principle of data observability is to gain visibility into the business impact of problems in the data environment. This could include observing customer experiences with portals and data applications.

Access control, security, and data governance ranks fifth among cloud concerns. TDWI research has tracked attitudes toward locating important business data in the cloud changing considerably in recent years, moving from strong hesitancy to a higher comfort level. Still, organizations need to ensure that only authorized users can access certain data and that security and governance procedures are updated to handle hybrid multicloud data environments. The challenge is to balance access controls and governance with the organization's need to gain value from data assets.

Modern tools are providing greater flexibility and sophistication in operationalizing constraints so that rather than be limited to blanket, one-size-fits-all enforcement, organizations can tune access control, security, and data governance to what users plan to do with the data, their geolocations, and other attributes.

Data Integration for Speed, New Workloads, and Complexity

Data integration is a broad category of practices, processes, and technologies for producing valuable, integrated data sets for user views and access as well as application needs. At a basic level, most data integration processes, such as data pipelines, begin with extracting or collecting data from multiple sources and loading it into a central repository or viewing it virtually through a single layer. Some data pipelines offer real-time data streaming to move data as fast as possible, often to a target data lake or other storage.

Data integration processes often include much more than just getting data from source to target, noting that targets include application databases, data warehouses, data lakes, spreadsheets, and BI or analytics platforms. Depending on requirements, they can include data profiling, quality and cleansing steps, data enrichment, data validation, and data governance.

Tools often enable organizations to monitor and spot problems in ETL, ELT, or data pipeline processes and either escalate notifications for human intervention or automatically solve the problems. Some tools also enable analysts to examine source data as it is coming through the pipeline, validate it, and determine how they want to transform it given their requirements.

Organizations typically have numerous best-of-breed tools for each of these steps or have an integrated toolset or set of cloud services. Many data platforms offer native connectors or have built-in capabilities that allow authenticated users to extract, replicate, or move data. Some solutions map data definitions and schemas between sources, APIs, and targets. Data catalogs are valuable for helping users locate data, understand how data is related to other data, and determine what data integration steps (such as cleansing) are required.

Some developers and administrators use open source tools rather than (or to complement) commercial solutions. Another growing option is to shop for predefined data integration services in cloud marketplaces and exchanges.

Data integration is central to preparing data sets to meet requirements—and therefore to maximizing the value of data. As we did earlier in this report with data management, we asked research participants what types of data integration technologies, patterns, or services their organizations are currently using or plan to use. Also as we did for data management, we asked whether organizations are using the most prominent data integration technologies for one or more of the following use cases: data science and AI/ML; BI, OLAP, dashboards, and business analytics; applications (e.g., operational or e-commerce); and customer or partner engagement or data sharing.

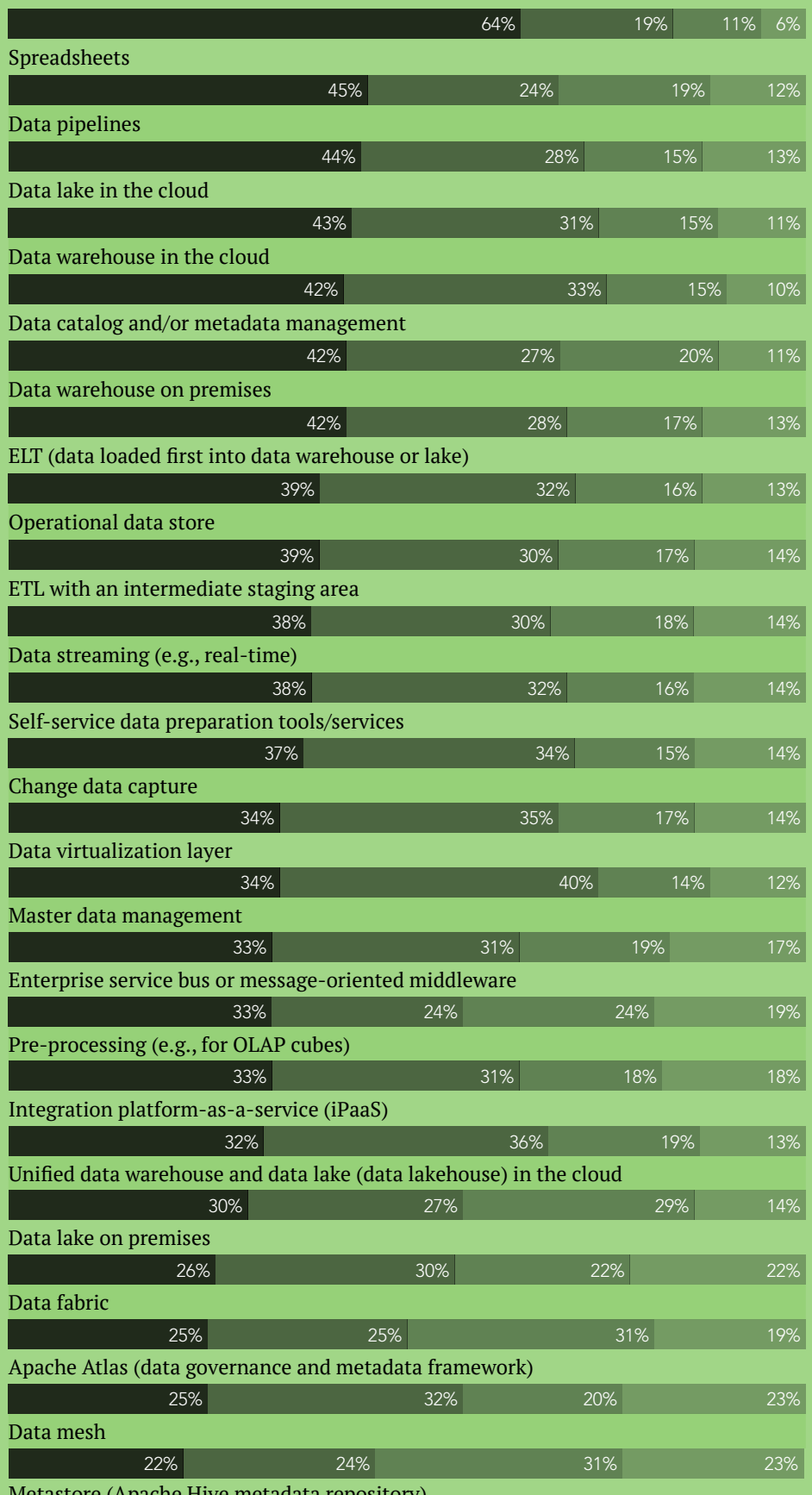
After spreadsheets, data pipelines are most common. Beyond ubiquitous spreadsheets (64% currently using), most organizations surveyed are currently using data pipelines (45%; 24% plan to use them; Figure 14). Just over one-quarter (27%) are using them for data science and AI/ML while 41% are using them for BI, OLAP, dashboards, and business analytics (figure not shown). Illustrating the variety of use cases for pipelines, 35% are using them for applications and 21% for customer or partner engagement or data sharing.

Data streaming solutions offer continuous flows through pipelines that gather and process data as soon as sources generate it. Users as well as AI/ML algorithms that drive data applications can then analyze streaming data and take a range of automated actions to serve use cases such as customer personalization, fraud detection, and real-time reporting. Some data integration tools and solutions automatically perform predefined data transformations and enrichment as needed on streaming data.

Figure 14 shows that 38% currently use data streaming and 30% plan to use it. Regarding use cases, our research finds that the highest percentage, 38%, use data streaming for BI, OLAP, dashboards, and business analytics; one-third (33%) use it for applications and 24% for data science and AI/ML (figures not shown). Nearly as many organizations surveyed (37%) currently use

Figure 14

Which of the following types of data integration technologies, patterns, or services are currently in use or planned for future use at your organization?



Based on answers from 325 respondents. Ordered by "currently using" responses.

change data capture (CDC) systems; 34% plan to use CDC. These systems capture changes to data and data structures continuously, typically landing updates in a data warehouse or BI/OLAP platform. Users then have access to the most current, near-real-time data with minimum disruption to operational and transaction systems.

> *Data streaming offers continuous flows for users as well as the algorithms that drive data applications; 38% currently use data streaming and 30% plan to use it.*

Organizations use ETL and ELT for a spectrum of use cases. Interestingly, our research shows a higher percentage of organizations surveyed (42%) are using extract, load, and transform (ELT) processes than traditional ETL (39%). With ELT, organizations build data pipelines to load (or replicate) data into a target repository before data transformation. ELT reduces some data movement by not requiring users to load the data first into a separate staging area and then, after transformation, into the target repository. Organizations use ELT to take advantage of the power and scale of cloud data processing platforms to run data transformations; this mode is frequently called in-database pushdown optimization.

However, despite ELT's growing popularity, ETL is still the appropriate choice for some workloads. An organization should make the decision based on its requirements. Some data integration solutions can automatically choose which mode to use. Our research finds that organizations are using ETL and ELT for a broad range of use cases, most prominently BI, OLAP, dashboards, and business analytics (47%, not shown). One-third (33%) are using ETL or ELT for applications, 28% for data science and AI/ML, and 22% for customer or partner engagement or data sharing.

Data virtualization, data fabrics, and data mesh show strong current and planned use. Just over one-third of organizations (34%) are currently using data virtualization and 35% plan to use it. Rather than depend on moving data to a central location such as a data warehouse or data lake, 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 in order to access them. Our use case research finds that the largest percentage of respondents (40%, not shown) are using data virtualization for BI, OLAP, dashboards, and business analytics; 18% are using it to support data science and AI/ML. Over one-quarter (27%) are using data virtualization for applications.

Significant percentages of organizations surveyed are currently using data fabric and data mesh frameworks and solutions to manage and govern highly distributed data environments such as hybrid multicloud. Although different, both approaches try to enable easier data flow among platforms, applications, and users while shielding users from underlying complexity. About one-quarter of research participants say their organizations are currently using a data fabric or data mesh (26% and 25%, respectively). Somewhat more plan to use one (30% and 32%, respectively). Data fabrics and data mesh will be discussed more later in this report.

Data lakes in the cloud are more common than those on premises. 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. As a central repository, the data lake allows organizations to let workloads dictate how to categorize, cleanse, and otherwise prepare data, particularly for analytics and AI/ML.

> *Data lakes are more common in the cloud now than on premises (44% versus 30%). Two in five organizations surveyed are using them for BI, OLAP, dashboards, and business analytics.*

The research shows that more organizations are taking advantage of affordable cloud storage to locate data lakes in the cloud; 44% currently have a data lake in the cloud compared to 30% on premises. Although data lakes are commonly associated with data science and AI/ML, in the research for this report, we find that more are using them for BI, OLAP, dashboards, and business analytics (40% compared to 28% for data science and AI/ML; 32% are using a data lake for applications, not shown). This suggests that some organizations are providing a broader range of users, not just data scientists, with direct access to the data lake.

Data warehousing in the cloud is about equal to data warehousing on premises. In 2021, TDWI research showed that 42% of organizations surveyed used an on-premises data warehouse and 33% had one located on a cloud platform.[4] About the same number (35%) said they were planning to have a cloud-based data warehouse. This year's results show that some of these plans have come to fruition. In Figure 14, 43% say they have a data warehouse in the cloud. The same percentage as in 2021 (42%) say they have a data warehouse on premises. Almost a third (31%) now plan to have a data warehouse in the cloud compared to 27% planning to deploy one on premises.

Data warehouses offer access to trusted, cleansed, summarized, and aggregated records organized for reporting and analysis across a variety of dimensions. With data warehousing to consolidate transformed and trusted data, users do not have to access each underlying source and know the way records at the sources are stored. A data warehouse offers historical data, but modern data warehousing processes can reduce latency to supply near-real-time, live data.

As with data lakes, organizations are taking advantage of scalable cloud storage and processing to go beyond the traditional fixed limitations of on-premises data warehouses. Regarding use cases, not surprisingly the largest percentage of respondents are using a data warehouse for BI, OLAP, dashboards, and business analytics (49%; not shown). However, 26% are using one to support data science and AI/ML. Just over one-third (34%) are using a data warehouse for applications, while 18% are using one for customer or partner engagement or data sharing, which could include support for services offered in a data marketplace or exchange.

Figure 14 shows considerable interest among respondents for unifying data warehouses and data lakes. Nearly one-third (32%) are currently using a unified data warehouse and data lake in the cloud and 36% plan to use one. Tighter integration would support data strategies for consolidating disparate data silos.

[4] Ibid., Figure 3.

Some organizations in our research are using a cloud data lake as a data ingestion and staging area for ELT processing to provide faster loading into the target data warehouse or analytics platform. TDWI finds that some organizations are interested in tightening data lake and data warehouse integration to make it easier for a range of users to access the contents of the data lake directly to build single views of all relevant data. The combined access allows users to gain contextual understanding from analysis of semi- and unstructured data to supplement reports based on traditional structured data.

Data Integration Modernization Priorities

The importance of data integration makes it critical to keep modernizing and evaluating options for current and future workload requirements. A wise strategy is to integrate and coordinate data integration modernization with data warehouse and data lake modernization to avoid gaps and take full advantage of scalable cloud processing and storage. Today's tools and platforms enable organizations to reduce data latency and shorten the path to value for all data assets.

In Figure 15, we examine how organizations rank their modernization objectives for data integration, including ETL, ELT, pipelines, and data virtualization. Improving data quality is a major theme throughout this report. Here, it is the top-ranked objective for most research participants (63% identify it as their top objective).

Echoing earlier findings, we see that organizations want to reduce costs and monitor cost drivers (e.g., maintenance). This objective ranks third.

Enabling reuse is an important strategy for reducing costs. Respondents say that increasing reuse and flexibility for faster workload development and deployment is their second-ranked objective (45% made it their second-ranked objective). As workloads increase and become more complex and diverse, organizations need to take advantage of tools and frameworks that support smart automation and reuse of ETL and ELT routines, trusted data sets, queries, models, and more.

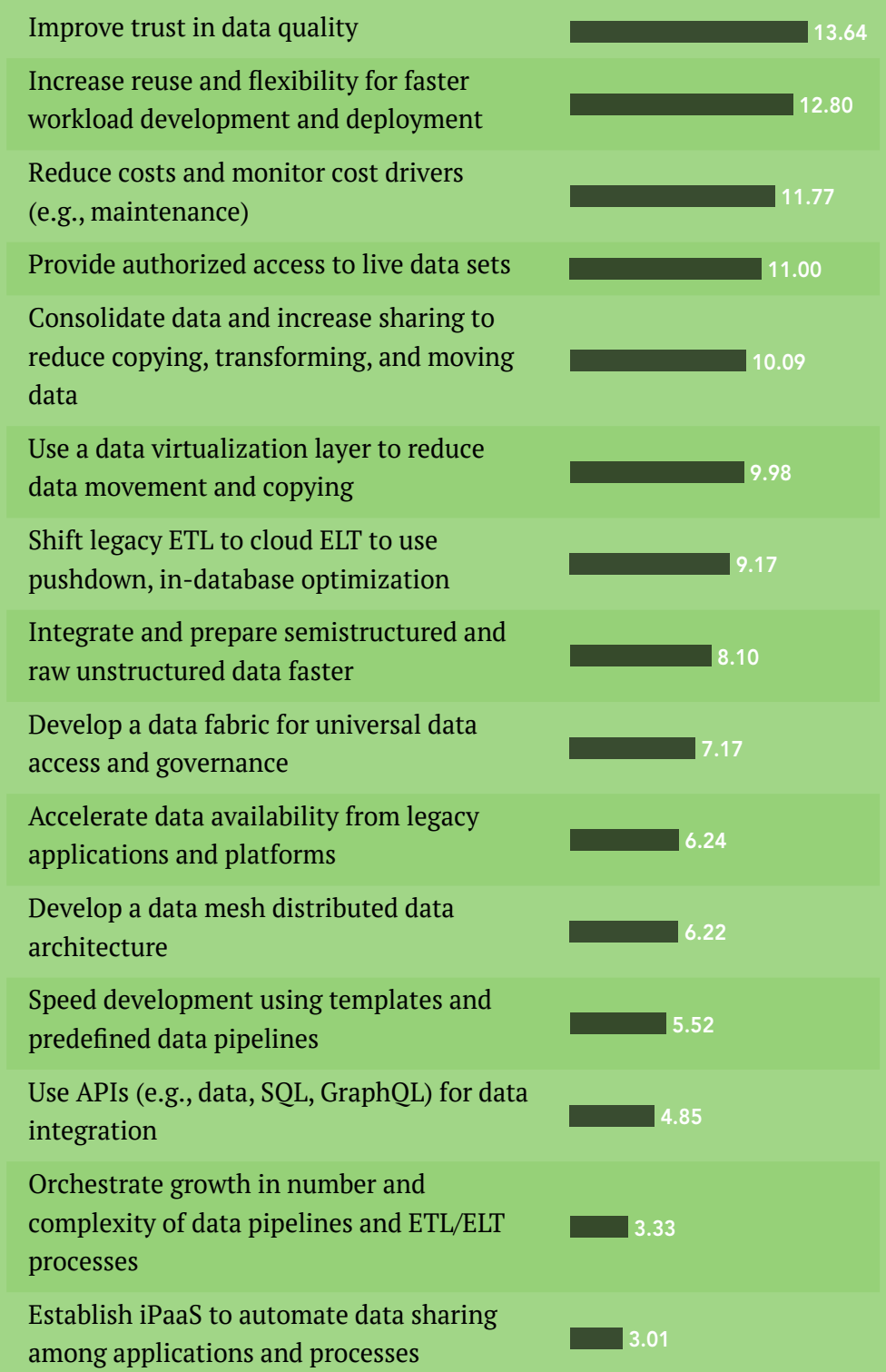
> *Organizations surveyed rank increasing data integration reuse and flexibility as one of their top objectives.*

Resources such as an enterprise data catalog can be helpful in centralizing knowledge about data and improving reuse of existing data sets, pipelines, and ETL/ELT routines. Organizations gain flexibility by having better documentation and consistency; it is easier to accommodate new requirements and workloads.

Providing authorized access to live data sets is a high priority. Research participants rank this objective fourth. Operational data applications may have numerous concurrent users or automated decision-making programs that need continuous updates about what has changed in the data. Business users may also need access to live data sets for operational reporting and analytics. Some organizations provide this by creating an operational data store or by using a zone in their data lake to provide access to near-real-time data. Organizations need to evaluate data quality requirements, however, because raw, live data is typically uncleansed and could have flaws and inconsistencies.

Figure 15

How would you rank the following modernization objectives for your organization regarding data integration, including ETL, ELT, pipelines, and data virtualization?



Based on answers from 316 respondents, who were asked to rank the options. Ordered by weighted average.

Consolidating data and increasing data sharing is a priority. Disparate data in numerous data silos burdens organizations with the overhead of too much data movement. The movement often comes with significant manual coding, which becomes part of the legacy "debt" mentioned earlier. Data silos increase the difficulty of providing users with single views of all relevant data; they also lead to inconsistent and incomplete governance. In Figure 15, consolidating data and increasing sharing to reduce copying, transforming, and moving data is ranked fifth.

Using data virtualization to reduce data movement ranks sixth. Another strategy for reducing the overhead caused by too much data copying, transformation, and movement is data virtualization. Some organizations use a data virtualization layer to provide faster access to live data sets at their sources rather than introducing the delays of moving and copying data to a target repository. Organizations surveyed made this the sixth-ranked objective.

Enterprise Data Catalogs and Semantic Layers

Data catalogs, business glossaries, and metadata repositories collect information about how data is defined and modeled, where it is located, and how models and schema may have changed. Thus, data catalogs can complement most data integration and pipeline processes, especially for distributed data strategies such as data virtualization, data fabrics, and data mesh.

> *Data catalogs are critical for gathering knowledge about data lineage. This is vital for data governance and overall data management. It also makes it easier for users to search for and find trusted data.*

Data catalogs are critical for gathering knowledge about data lineage, i.e., the data's origin and what has happened to it during its life cycle. This includes how organizations transform, replicate, and share data. This knowledge is vital for data governance and overall data management. Access to data lineage information also makes it easier for self-service business users to search for and find relevant and trusted data from diverse sources.

Data stewards play an important role in accelerating paths to business value. AI-infused tool automation in modern enterprise data catalogs helps data stewards, who have an important role in data governance. Data stewards oversee data quality processes, manage metadata and master data, and guide users to data that is fit for purpose. Organizations often struggle to identify data stewards, most of whom have "day jobs" as business subject matter experts (SMEs). Data catalogs can help organizations locate potential data stewards by looking at information collected about their data activity, knowledge of the data and data ownership, and understanding of relevant data governance constraints. Tools in some data catalog solutions can automatically attach certain data sets and data integration processes to newly identified data stewards.

Tool automation enables organizations to scale stewards' data governance oversight and have better visibility into data quality, classification, and documentation issues. When data sources change, an AI-infused data catalog could automatically inform data stewards, who could then ensure that dependent users and applications are informed and prepared. In addition, AI-driven automation enables data catalogs to uncover patterns and data relationships to provide recommendations for analytics. This can accelerate model development and increase data consistency across models.

Data catalogs fit with several other technologies and practices such as master data management (MDM) to form the semantic layer. By collecting and standardizing knowledge about business and data definitions, semantic layer technologies have long played a critical role in ensuring the quality and consistency of business representations of data for reporting, calculations, and multidimensional modeling. BI/OLAP systems and many business applications typically have a semantic layer to support self-service analytics and sharing of calculations, data set selection, search, and workflows. However, a challenge many organizations face is that different semantic layers are often not consistent or easily integrated.

In Figure 14, about two in five organizations (42%) surveyed currently use a data catalog and/or metadata management system, significantly more than the 31% our 2021 research indicated.[5] A year ago, 36% said they were planning to use such a system. Along with the rise in current usage, in this year's research, 33% say they plan to use a data catalog and/or metadata management system.

We also see in Figure 14 current usage rates for two prevalent Apache open source metadata technologies. One-quarter of respondents (25%) say they are currently using Apache Atlas, a metadata management and data governance framework primarily for Apache Hadoop environments. Another 25% plan to use it.

Nearly as many (22%) are implementing Apache Metastore, one of the main components of the Apache Hive data warehouse architecture, which is part of the larger Hadoop environment. Although technologies in the Hadoop environment have evolved considerably away from MapReduce and other older systems toward Apache Spark and cloud-native services, many newer tools still support organizations' use of Metastore as a data catalog for data lakes.

Top Five Semantic Layer Priorities

Organizations are interested in using semantic layer technologies to address a range of issues important to user satisfaction in accessing and viewing trusted data and meeting enterprise priorities for data governance, integration, and management. In our research, organizations indicate that the following are the top five priorities for a semantic layer. We elaborate on this analysis with the data shown in Figure 16 about relevant satisfaction levels for data catalogs and other metadata management.

> *Enabling users to access data using common business terms and hiding technical complexity are the top two semantic layer priorities.*

[5] Ibid.

#1: Enable users to access data using common business terms. This priority includes finding and accessing data across distributed sources. Organizations use semantic layer technologies to enable faster and better integration of data from silos and external sources so users do not face delays and complexities in finding and accessing the data.

More than half of research participants (58%) indicate that their organizations are satisfied with their data catalog or other metadata management for making it easier for users to search for and find data. More than half (54%) also say their data catalog or other metadata management improves user collaboration in choosing data sets, which is more challenging in distributed data environments.

#2: Hide technical complexity of data access from business users. Organizations want semantic layers to be part of the abstraction layer that enables nontechnical users to interact with data without technical skills. Semantic layers should help deliver unified, consolidated views of data across the organization. Semantic layers are a critical component in data virtualization; 48% of organizations surveyed are satisfied with the integration of the data catalog with their data virtualization layer.

Data catalogs and other metadata management enable the majority of organizations surveyed to gain a more complete inventory of data assets with less technical complexity. Over half of organizations (55%) are satisfied with their systems for this purpose as well as providing documentation of objects available for user consumption. This information about the data is helpful in developing data pipelines as well as transformation and preparation routines; 55% are satisfied with their data catalogs and metadata management for these jobs.

#3: Refine business representations of data to be accurate and usable. Semantic layer technologies are important for interfacing business requirements and context with underlying data. Semantic layers must support the creation of calculated fields, dimensions, and aggregations and enable easier discovery of data relationships important to business analytics.

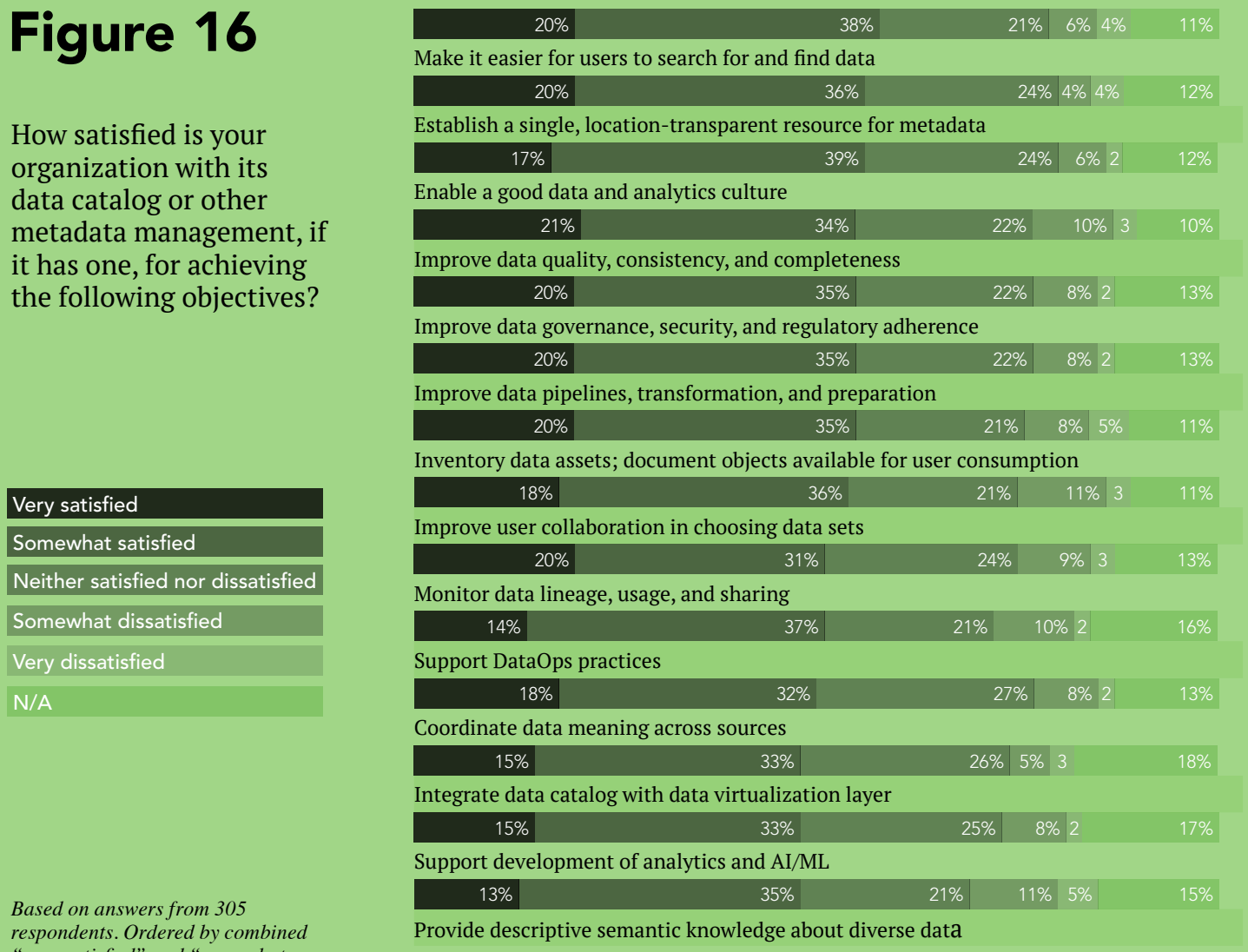
Contextual knowledge can make data quality improvement faster and more consistent. More than half of organizations surveyed (55%) say that data catalogs as part of semantic layers improve their data quality, consistency, and completeness.

#4: Map complex data to business terms (e.g., product, customer, revenue). Business glossaries, MDM, and data catalogs all are important for addressing this priority. Accurate mapping that can be modified over time helps identify data relationships in the context of business requirements and data governance. Our research suggests that there is room for improvement in using data catalogs and metadata management for providing descriptive semantic knowledge about diverse data; 48% are satisfied, but only 13% are very satisfied.

#5: Provide common definitions for metrics, dimensions, and other objects. This priority calls on the semantic layer to eliminate all-too-common confusion when users work with different BI/analytics tools and applications. Enterprise data catalogs as part of semantic layers can play an important role. More than half of organizations surveyed (56%) say they are satisfied with their data catalog and metadata management for establishing a single, location-transparent resource for metadata and 50% are satisfied with using the system to coordinate data meaning across sources.

Figure 16

How satisfied is your organization with its data catalog or other metadata management, if it has one, for achieving the following objectives?



Based on answers from 305 respondents. Ordered by combined "very satisfied" and "somewhat satisfied" responses.

Finally, slightly more than half of organizations (55%) are satisfied with data catalogs and metadata management for data governance, security, and regulatory adherence. One of the critical aspects of data governance is monitoring data lineage, usage, and sharing. Somewhat fewer (51%) are satisfied with this, indicating room for improvement.

Distributed Data: Data Fabric and Data Mesh Strategies

Distributed data environments with many disparate data silos present numerous challenges. Users become frustrated when they cannot access all the data they need when they need it without skilled experts to code programs for data integration. Data governance is difficult and inconsistent due to poor visibility across silos.

Many organizations have a strategy of consolidating data onto cloud platforms, but it is likely that data silos will continue to exist. As noted, hybrid multicloud environments are common and present distributed data management challenges, including additional data silos.

> *Traditional divisions between operational applications and analytics are breaking down as demand grows for data-hungry, AI-driven applications.*

Organizations need new strategies for handling distributed data as traditional divisions between operational applications and analytics break down under pressure from development of data-hungry, AI-driven data applications. Technologies such as in-memory computing, scalable unified data platforms, and automated analytics are enabling organizations to reduce latency and provide a broader range of users and applications with near-real-time data insights.

Reduced latency is vital for use cases such as financial trading systems, e-commerce and customer personalization, manufacturing, utility grid monitoring, and other business activities. However, if organizations cannot overcome problems in managing distributed data environments, they will face difficult obstacles in reducing latency and supporting near-real-time, data-driven business strategies.

Organizations are seeking seamless and unified data flows and coordinated analytics. The changing landscape is increasing interest in data fabrics and data mesh architectures. Although they differ, the two approaches share the goal of a universal approach to integrating diverse components of physically distributed data environments. Semantic layers consisting of metadata catalogs, MDM, knowledge graphs, ontologies, and more are important to data fabrics and data mesh to enable easier location of relevant data and effective and less-onerous data governance.

Data fabrics use semantic knowledge and virtualization. A data fabric is a framework (or "design concept," as some call it). It runs on a virtualization layer and uses knowledge contained in metadata assets and semantic layers to make it easier to develop and operationalize analytics across a distributed data environment. Data fabrics can use knowledge graphs and graph databases to make it easier and faster to discover data relationships in complex, distributed data and store records of the relationships for analytics.

Data virtualization is important for accessing and governing data in a data fabric. Integrated with semantic layer resources such as a data catalog, a data fabric can ensure that users' selected data sets are relevant and sensible. Data integration and data governance services enable trusted data to

flow easily. Users and applications do not have to use specialized code to access each data silo. The organization can use the data fabric as a logical layer for controlling access and meeting defensive data governance requirements for sensitive data protection.

> *Data virtualization, integrated with the semantic layer, is important for accessing and governing data in a data fabric.*

Data meshes enhance the business value of decentralized data architecture. Similar to a data fabric, a data mesh takes advantage of centralized data governance and standardized data integration and interoperability. However, a data mesh takes a different focus. Rather than emphasize centralized control over a distributed architecture, a data mesh leaves more control with decentralized domains. It seeks to increase the potential of self-service data infrastructure and federated computational governance.

The data mesh approach takes an outcome-oriented view, treating data as a product that has a purpose in solving problems and improving business outcomes. Independent data producer teams have the freedom to increase the value of data as they see fit. The architecture views domains as independent and gives teams more control. The data mesh limits centralized coordination to essentials such as data governance.

Distributed Data Solution Drivers

TDWI asked research participants to rank the importance of a series of key objectives to their organizations for establishing a unified distributed data architecture such as a data fabric or data mesh framework (see Figure 17). Not surprisingly, the top-ranked objective highlights the basic priority for gaining business value from data: to simplify and accelerate user access to distributed data (55% rank it first).

Data fabrics and data meshes enable organizations to meet this top objective. Ranked third is a related objective that points to the importance of underlying data virtualization to data fabrics: creating a logical layer for a single point of access over disconnected silos. Many organizations additionally want to reduce unnecessary data movement, which can increase delays and costs and expose data to risks of unauthorized access. This ranks fifth among objectives.

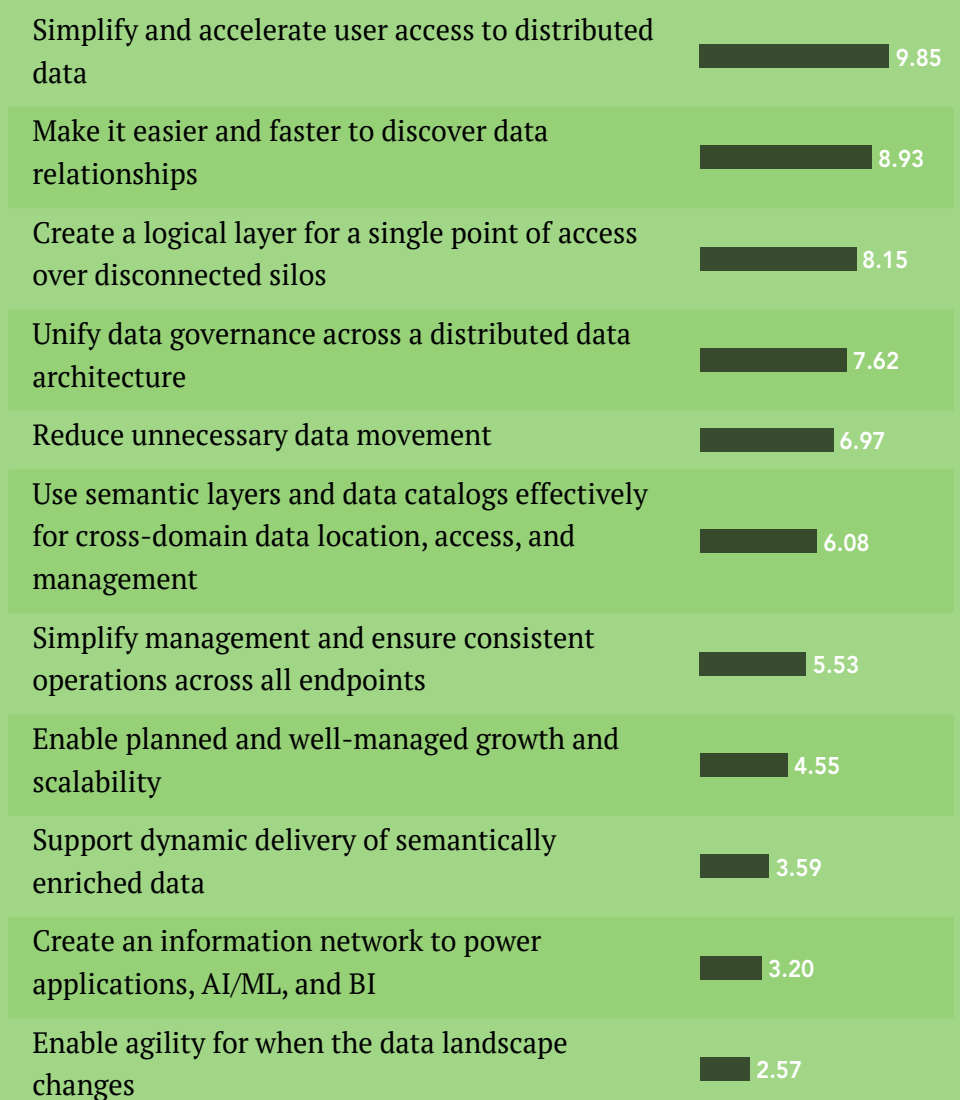
Making it easier and faster to discover data relationships ranks as the second-highest objective. This demonstrates the importance of the semantic layer technologies mentioned above for accelerating analytics in a distributed data environment. A related objective, using semantic layers and data catalogs effectively for cross-domain data location, access, and management, ranks sixth.

> *Unifying data governance across a distributed data architecture ranks as the fourth-highest objective among organizations.*

Unifying data governance across a distributed data architecture ranks as the fourth-highest objective. As noted, data governance is a common pain point for organizations that have highly distributed data, including in hybrid multicloud data environments. Data fabric and data mesh approaches both elevate the importance of consistent and continuous data governance.

Figure 17

How would you rank the importance of the following objectives to your organization for establishing a unified distributed data architecture, such as a data fabric or data mesh framework?



Based on answers from 310 respondents, who were asked to rank the options. Ordered by weighted average.

Organizations should evaluate whether their current and future data environment aligns with either the data fabric or data mesh approach—or both: organizations could use them in a complementary fashion. Both options include defensive data governance strategies important to protecting sensitive data in accordance with data privacy and other regulations and internal rules. They also afford implementation of offensive data governance priorities for data curation to increase quality and value of data assets for users. Within a data fabric or data mesh, modern data governance can take advantage of AI-driven automation to enable organizations to address all data governance priorities.

Recommendations

To conclude this report, here are 10 best practices for maximizing the value of data platforms as well as independent data integration and data management technologies and services. Our research shows that these 10 are key for meeting current and future demands and challenges.

Have a plan for hybrid multicloud data environments. Successful cloud migration is a top priority for organizations participating in our report. They want the scalability, flexibility, cost elasticity, and speed of deployment that cloud computing offers. However, organizations that have significant on-premises data will have hybrid multicloud environments for some time to come while they migrate to the cloud in phases or plan to keep some data assets on premises permanently.

Rather than have piecemeal cloud migration that makes data silos out of important data assets, develop a plan for hybrid multicloud. Evaluate distributed data frameworks such as the data fabric and data mesh to establish more unified and holistic data access, management, and governance.

Increase the value of data assets by expanding data sharing. In most cases, data becomes more valuable when data owners share data. Complete views of customer behavior and buying patterns, for example, can enable internal marketing, merchandising, and other departments (as well as external partners) to gain richer data insights. Confident data sharing requires data governance to protect sensitive data from unauthorized access and to improve data curation so data quality, completeness, and consistency meet requirements.

Semantic layer technologies play an important role in enabling faster location of data and accurate understanding of the data's relationship to business entities. Cloud data platforms provide required scalability and can make data more easily available. Data sharing should be an important part of your data strategy.

Move beyond legacy strategies to address AI-infused data application needs. To compete, organizations need operations to benefit from new, data-rich applications. These increasingly depend on analytics, including AI/ML, to monitor conditions, detect patterns, drive real-time automated decisions, and guide human intervention. Traditional divisions between BI and analytics on one side and transactional applications on the other hold organizations back.

Organizations need to evaluate NoSQL systems for supporting modern data applications. Determine your data application development and deployment requirements and how they affect plans for data platforms that feature data integration and data management capabilities as well as independent data integration and management technologies and services.

Strengthen your semantic layer with a data catalog and related technologies. The semantic layer is critical to enabling users to locate and make sense of voluminous, diverse, and distributed data as it relates to their business context. Organizations benefit from semantic layer technologies, particularly an enterprise data catalog, to govern data inventories, monitor data life cycles and lineage, and know who uses and shares the data.

Research for this report shows that organizations gain value from a robust semantic layer that could include an enterprise data catalog, data lineage tools, MDM, and knowledge graphs. Evaluate

modern, AI-infused technologies for establishing and improving these shared resources.

Unify distributed data environments with data virtualization, data mesh, and data fabric. As data environments expand, many organizations suffer from limited and siloed data access, incomplete data governance, and difficulty integrating data among critical data-driven applications. Research in this report shows strong interest in establishing unity on top of distributed data environments, including hybrid multicloud.

Although data migration and consolidation on scalable cloud platforms addresses some challenges, organizations will need holistic strategies to encompass their distributed data environments. Evaluate data virtualization, data mesh, and data fabric technologies and frameworks for relevance to your data strategy.

Plan for continued data democratization and self-service. Users are excited about opportunities to use data and develop analytics to improve business insights and outcomes. They need trusted, relevant data assets and the ability to view and access the right data at the right time. Latency and bottlenecks in data integration, transformation, and preparation are sources of frustration.

Achieving a balance between self-service analytics and enterprise resources and governance is important. Organizations need to observe users to learn what satisfies them or stalls their progress and innovation with data. Do not overlook opportunities for AI augmentation to drive automation, predefined queries and data set selection, and recommendations that could help users produce faster insights.

Reduce data latency and support real-time analytics. Our research shows strong interest in fresher data, quicker updates, and live data access. Emerging data applications depend on timely data and continuous updates and insight availability. Use data intelligence and observability to identify bottlenecks, delays, and disconnects in data integration and data pipelines and throughout data life cycles in both traditional data systems and AI-infused data applications.

Ensure that shared enterprise data systems such as data catalogs are available and easy to use to reduce delays in locating relevant data. Evaluate data fabric and data mesh approaches for how they reduce data latency and enable progress toward real-time data access and insights.

Improve data trust with modern data governance. This research shows that improved data governance is a top priority as organizations democratize data and analytics, develop data applications, and seek to improve data sharing. Without data governance, organizations cannot overcome challenges that arise when they increase data volume, speed, and variety and expand workloads.

One common concern is that data governance involves significant manual work and administration, so users view data governance as a burden rather than a benefit. Advances in AI/ML-infused automation enable organizations to use data intelligence effectively to improve data governance and move beyond manual work and administration. Modernize data governance and make it core to your data strategy.

Prioritize support for analytics and AI/ML development and operationalization. Our research shows that advancing with visualization and analytics is a top objective, including through development of projects that apply AI/ML, NLP, and text analytics. Migrating to modern cloud data

platforms and distributed data frameworks and technologies is important for ensuring continuous data access, scalability, and flexibility for analytics and AI/ML.

Data scientists need access to both the cloud data warehouse and data lake as well as distributed data through data virtualization, data fabrics, and data mesh. As data applications grow, address developers' needs to use NoSQL and languages such as Python, R, and Scala, not just SQL. Analytics and AI/ML-infused applications increasingly need real-time data access and faster data integration.

Develop a data strategy to overcome complexity and delays. Having a solid and forward-looking data strategy will help your organization make thoughtful decisions involving cloud data migration and adoption of new data platforms. A data strategy will help you prioritize modernization of data integration and management capabilities inside data platforms as well as independent technologies and services. A data strategy should connect business priorities with the data environment so business issues are the context for modernization and cloud migration.

Data strategy is also important to building a strong data and analytics culture in the organization. Committees of business and IT stakeholders, potentially led by a CDO, should meet regularly to address users' challenges, keep abreast of data governance priorities, and articulate future plans.

mongodb.com

MongoDB is a developer data platform company empowering innovators to create, transform, and disrupt industries by unleashing the power of software and data. Headquartered in New York, MongoDB has more than 35,000 customers in over 100 countries. The MongoDB database platform has been downloaded over 265 million times and there have been more than 1.5 million registrations for MongoDB University courses.

Learn more at mongodb.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.

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

E info@tdwi.org

tdwi.org