research

BEST PRACTICES REPORT

Q4 2021

Modernizing Data and Information Integration for Business Innovation

How organizations can use new solutions and practices to meet expanding data demands

By David Stodder

tdwi
Transforming Data With Intelligence™

Research Sponsors

Actian

Denodo

SnapLogic

Stardog

Trifacta

Modernizing Data and Information Integration for Business Innovation

How organizations can use new solutions and practices to meet expanding data demands

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 insight and best practices for organizations implementing BI, analytics, performance management, data discovery, data visualization, and related technologies and methods. He is the author of TDWI Best Practices Reports and Checklist Reports on AI for BI, cloud analytics, data visualization, customer analytics, big data analytics, design thinking, and data governance. He has chaired TDWI conferences on visual BI and analytics, business agility, and analytics. Stodder has provided thought leadership on BI, information management, and IT management for over two decades. He is an industry analyst, having served as vice president and research director with Ventana Research, and he was the founding chief editor of *Intelligent Enterprise* and *Database Programming & Design*. You can reach him at dstodder@tdwi.org, @dbstodder on Twitter, and on LinkedIn at linkedin.com/in/davidstodder.

About TDWI

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.

About the TDWI Best Practices Reports Series

This series is designed to educate technical and business professionals about new business intelligence, analytics, AI, and data management technologies, concepts, or approaches that address a significant problem or issue. Research is conducted via interviews with industry experts and leading-edge user companies and is supplemented by surveys of business and IT professionals. To support the program, TDWI seeks vendors that collectively wish to evangelize a new approach to solving problems or an emerging business and technology discipline. By banding together, sponsors can validate a new market niche and educate organizations about alternative solutions to critical problems or issues. To suggest a topic that meets these requirements, please contact TDWI senior research directors Fern Halper (fhalper@tdwi.org), James Kobielus (jkobielus@tdwi.org), and David Stodder (dstodder@tdwi.org).

Acknowledgments

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

Sponsors

Actian, Denodo, SnapLogic, StarDog, and Trifacta sponsored the research and writing of this report.

Research Methodology and Demographics

Report purpose. Data and higher-level information integration processes are critical to how organizations gain new business insights, understand data relationships, collaborate internally and externally, improve efficiency, and drive data-driven innovation. Organizations need to modernize data and information integration to handle the data demands of digital transformation, self-service visualization, and analytics democratization. This report examines current challenges, solution options, and strategies for the future.

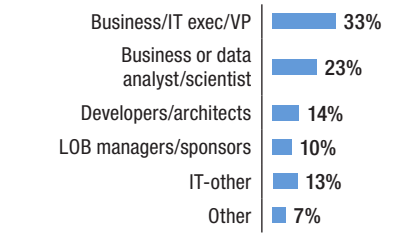
Survey methodology. In August and September 2021, TDWI sent invitations via email to business and IT professionals in our database, asking them to complete an internet-based survey. The invitation was also posted online and in publications from TDWI and other firms. The survey collected responses from 437 respondents. Not all respondents completed every question; however, 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 telephone interviews with IT and business executives and managers, technical users, and BI, analytics, AI, and data management experts. TDWI also received briefings from vendors that offer products related to the topics addressed in this report.

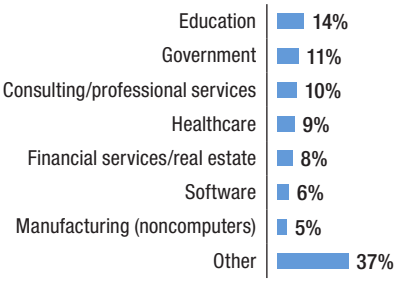
Survey demographics. One-third of respondents are business or IT executives and VPs (33%). The second-largest percentage consists of business or data analysts and data scientists (23%); third largest is developers and data, application, or enterprise architects (14%). Line-of-business (LOB) managers and business sponsors account for 10% of the respondent population. "IT-other" and "other" titles account for one-fifth of the total (20%).

Education is the largest industry group (14%), followed by government (11%), consulting and professional services (10%), healthcare (9%), financial services and real estate (8%), and software (6%). Respondents from manufacturing (noncomputers) accounted for 5%, followed by insurance (4%) and retail, CPG, wholesale, and distribution (4%). Construction and engineering and nonprofits and trade associations each accounted for 3%. The remaining 23% are from a variety of other industries. Just over four in five survey respondents reside in the U.S. (81%), with Asia and Pacific Islands (5%), Canada (5%), and other regions following. Respondents come from enterprises of all sizes.

Position

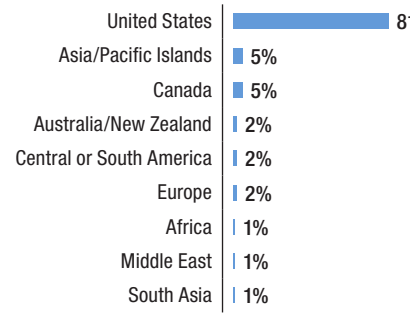


Industry



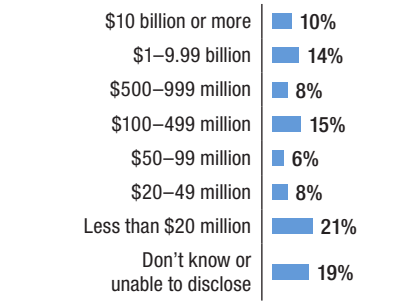
("Other" consists of multiple industries, each represented by 4% of respondents or less.)

Geography



Number of Employees

Company Size by Revenue



Demographics based on 437 respondents.

Executive Summary

Getting complete and quality views of data about subjects of interest is essential to nearly every business decision or action today.

Data and higher-level information integration are the foundation for reaching data-driven objectives. Getting complete, quality views of data about subjects of interest is essential to nearly every business decision or action. For analytics and data exploration, people across enterprises need timely access to new and diverse data that is integrated, transformed, and managed, not in a one-size-fits-all fashion, but flexibly to fit diverse types of data, knowledge requirements, and workloads. This puts pressure on organizations to modernize data and information integration, including by taking advantage of advances in artificial intelligence (AI) and automation embedded in tools and technologies. Organizations also increasingly need cloud services that supply powerful, modern computational processing for higher speed, scale, and agility.

This TDWI Best Practices Report focuses on strategies for solving data and information integration challenges. Organizations want to increase user satisfaction, support flexibility and innovation, and enable people to collaborate effectively to achieve business objectives such as operational efficiency, richer customer engagement, growth, resilience, and risk mitigation. The report examines business and technology trends and drivers; we analyze research data about challenges in using current solutions and explore where organizations need to modernize.

The report finds that organizations need to address interrelated issues to advance with data and information integration. They need to support expanding data visualization and analytics performed by data-savvy business users who want self-service power and flexibility for data access, blending, and preparation—and want to make decisions based on deeper knowledge of data relationships.

At the same time, organizations are looking for solutions, especially in the cloud, that support self-service expansion but offer cost efficiency, governance, and fewer headaches due to data fragmentation. This report finds that analytics as well as AI and machine learning (AI/ML) development are important drivers behind modernization. Organizations need smart automation to reduce latency in data pipelines and integration processes so they can provision fresher or real-time data for analytics, AI/ML, operational dashboards, and embedded intelligence.

Organizations face challenges stemming from disparate data silos, confusion about data definitions and master data, and difficulty discovering and analyzing complex data relationships. Traditional data and information integration systems, designed for a defined set of business processes, can limit the value people derive from new data types that do not fit easily into standard samples and aggregations. This report looks at common pain points and discusses strategies for solving them and moving beyond legacy constraints.

Relevant solutions discussed include data virtualization, data catalogs, knowledge graphs, master data management, data fabrics, and data silo consolidation into unified cloud data management. Rather than competing, solutions often complement each other within a larger data strategy to enhance agility so that organizations can handle both routine and unanticipated, ad hoc workloads. Building deeper knowledge about the data including its location, lineage, and relationships not only helps people find data faster and gain comprehensive views; the knowledge is critical to data governance and adherence to regulations.

This TDWI Best Practices Report offers recommendations that highlight priorities for modernizing data and information integration for business advantage.

Business Demands Drive Modernization

Data and information integration ultimately serve business objectives. To gain actionable insights into customer behavior, competitive pricing, population health, the resilience of supply chains, and many other key issues, decision makers want complete and integrated views of data. Executives, managers, and frontline personnel need collections of timely, quality, and trusted data for informed collaboration. The success of initiatives such as the digital transformation of business processes rests on how fully organizations realize value from new and diverse data sources for analytics, operational efficiency, and business integration.

Just as business is a combination of the routine and the unexpected, data and information integration must support both repeatable requirements and unexpected, ad hoc data demands driven by immediate needs. To empower a spectrum of users to work effectively with data, integration tools and services must be easy to use. Solutions need to relieve all users of time-consuming manual data integration and data chaos caused by poor, ill-defined, and inconsistent data. Organizations require solutions that support data governance policies and close observation of adherence to data privacy and industry-specific data use regulations.

Just as business is a combination of the routine and the unexpected, data and information integration must support both repeatable requirements and unexpected, ad hoc data demands driven by immediate needs.

This TDWI Best Practices Report examines data integration and higher-level information integration based on metadata, master data, and semantic data integration. Technologies in this space are evolving from significant hand coding and limited, often brittle, programs to scalable, agile, increasingly automated, and AI-infused solutions. As data architectures expand to cover a variety of distributed platforms and user interactions with data through different channels (for example, mobile devices, web portals, and desktop computers), integration technologies are broadening into fabrics that provision users with data views and access to the entire environment, not just single, highly structured data warehouses.

In this report, we explore the current state of organizations' data and information integration as well as top challenges and future plans. The report discusses technology innovations that offer opportunities to overcome limitations and delays so that even as the data universe expands, users have accurate, timely, and relevant data views and access. Innovations are occurring in a variety of technologies including data catalogs and metadata repositories, semantic integration and knowledge graphs, data transformation, data virtualization, data pipelines, data lakes and data warehouses, and data fabrics. Cloud computing is changing the landscape for all technologies and practices; this report examines how data and information integration modernization affects cloud migration.

Modernization Objectives

To begin, we asked research participants about their most important objectives for modernizing data integration and management (see Figure 1). Just over half (51%) indicate that a top objective is data democratization, i.e., supporting expansion in data visualization, business intelligence (BI), and analytics to serve more users. Nearly a third (29%) want to modernize so they can provision single, complete views of the truth, which is typically one of the primary requirements of users regarding data integration.

Just over half of research participants indicate that data democratization is a top modernization objective; nearly a third want to provision single, complete views of the truth.

Organizations show strong interest in upgrading data integration and management to drive analytics workloads, with 43% indicating that increasing data science, predictive modeling, and AI/ML is a top objective. The advantage organizations can gain from real-time data integration and analytics is a modernization driver for 40% of respondents. Using advanced analytics to spot trends and enable proactive response is critical to reducing business, market, financial, and fraud risk; reducing such risks is a top objective of 29% of respondents.

Which of the following are currently the most important objectives for modernizing your organization's data integration and management? Select up to five.

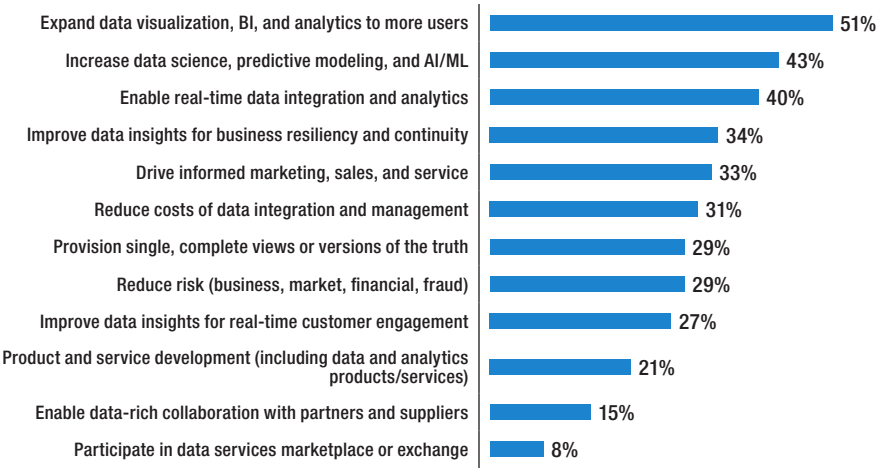


Figure 1. *Based on answers from 437 respondents.*

We can see important business drivers behind the modernization of data integration and management in several other objectives chosen by significant percentages of respondents:

- Over a third (34%) want to improve data insights for business resiliency and continuity—a major challenge given the disruptions spawned by the COVID-19 pandemic and recent supply chain breakdowns affecting numerous industries as well as consumer markets
- One-third (33%) say driving more informed marketing, sales, and service is a top objective; organizations prize accuracy and insight for personalization, greater efficiency, and 360-degree views of customer interaction
- Over one-fourth (27%) are seeking to improve data insights for real-time customer engagement
- More than one in five (21%) want to modernize to fuel product and service development, including data and analytics products and services (i.e., monetization)

Additional research participants highlight reducing costs and enabling better collaboration on data, including through participation in data services marketplaces and exchanges. This report discusses these topics later.

Satisfaction for Critical Applications and Services

Organizations need to curate and provision essential data for an ever-growing variety of applications and services.

Organizations need data integration and management systems to curate and provision essential data for a growing variety of applications and services that users interact with on desktops, mobile devices, large-format screens in shared areas, and portals over the web. TDWI asked research participants about their organizations' satisfaction with current data integration and management for key applications and services.

Here are those applications and services for which organizations have the highest levels of satisfaction (figure not shown):

- **Spreadsheets.** One-third (34%) are very satisfied and 47% are somewhat satisfied but need some improvement. Spreadsheets remain a common tool for individual users to extract data from sources and organize, categorize, manipulate, and analyze it.
- **Enterprise BI and reporting and performance management.** Just 20% are very satisfied; 58% are somewhat satisfied, citing the need for some improvement. Often well established and tied to traditional data warehouses, enterprise BI and reporting systems are workhorses. However, most specialize in offering selected aggregations of highly structured and transformed historical data. About the same percentages of respondents are very and somewhat satisfied with data integration and management for performance management, including scorecards, key performance indicators (KPIs), and metrics. Somewhat fewer are satisfied with data provisioning for online analytical processing (OLAP); just 13% are very satisfied and 47% are somewhat satisfied.
- **Self-service dashboards and data discovery.** Fifteen percent are very satisfied and 56% are somewhat satisfied; 23% are unsatisfied and need a major upgrade. As users gain experience and expand beyond simple data consumption into deeper analytics, they place more intensive demands on underlying data integration and management.

Respondents indicate lower levels of satisfaction with data integration and management for analytics, including AI/ML, data catalogs, and data lineage:

- **AI/ML development and AutoML.** This use case garnered the lowest satisfaction levels, with 12% very satisfied, 33% somewhat satisfied, and 30% unsatisfied and looking for a major upgrade (25% don't know). The research results suggest that organizations investing in AI/ML and automated machine learning should focus on modernizing data pipelines for speed and scalability and streamlining access to raw, semistructured, and unstructured data as well as prepared data aggregations.
- **Understanding and monitoring data lineage.** Only 14% are very satisfied and 35% are somewhat satisfied; 35% are unsatisfied and looking for a major upgrade (16% don't know). This report will discuss data lineage later in the context of data catalogs and data governance.
- **Data catalogs.** Twelve percent are very satisfied with their use and 37% are somewhat satisfied and want some improvement; 30% are interested in a major upgrade and 21% don't know. Later, we will discuss data catalogs and their integration with data pipelines, data virtualization, and data management.

Hindrances to Data-Informed Decisions and Realizing Value

In what areas are experiences with data integration and management falling short? We asked research participants which factors present the most significant hindrances to users making data-informed decisions and realizing value from data assets. The answers help clarify where organizations need to focus modernization efforts. In Figure 2, we can see that about three in five respondents (61%) find their users are stymied by data quality, completeness, and consistency issues. This result points to the importance of modernizing data integration and management to provision higher quality data.

About three in five respondents (61%) say data quality, completeness, and consistency challenges are the most significant hindrances to satisfactory data integration and management.

Regarding data integration and management, which of the following factors present the most significant hindrances to users making data-informed decisions and realizing value from data assets? Select up to five.

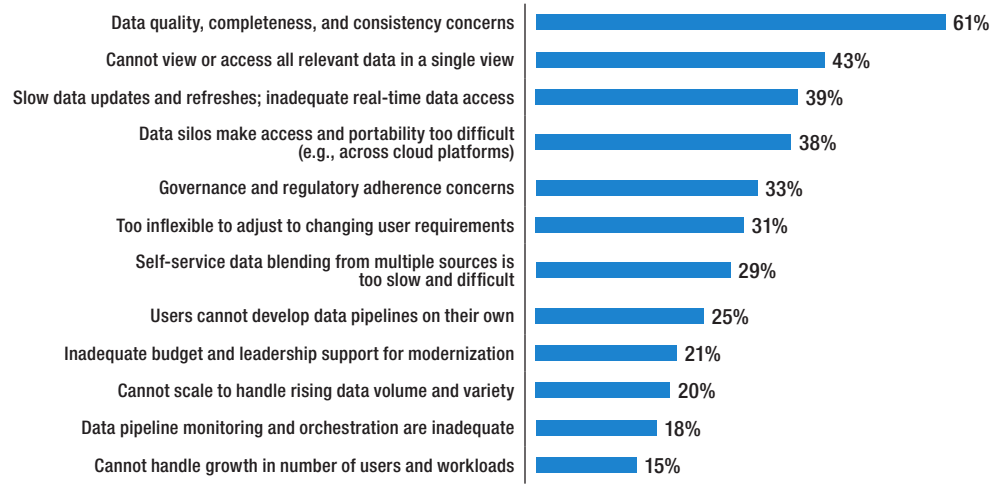


Figure 2. *Based on answers from 434 respondents.*

We noted earlier that gaining single views of data is a top objective. In Figure 2, we can see the need for improvement, with 43% noting that the inability to view or access all relevant data in a single view remains a hindrance. Nearly as many (38%) say data silos make access and portability too difficult, including across multiple cloud providers' platforms. Data silos are a hindrance to both data quality and gaining complete views of data. Governance and regulatory adherence concerns, which are more difficult to solve when many data silos exist, is a hindrance for 33% of respondents.

Percentages for other factors in Figure 2 suggest that organizations are facing agility challenges and are encountering difficulties responding to self-service needs. Nearly one-third (31%) say their current data integration and management is too inflexible to adjust to changing user requirements. Almost as many (29%) say self-service data blending from multiple sources is too slow and difficult to support users making data-informed decisions and realizing value from data assets. One-quarter (25%) indicate that users cannot develop data pipelines on their own.

We will discuss these and other hindrances shown in Figure 2—as well as the technologies important to solving them—as we move through this report's research. We offer best practices recommendations at its conclusion.

Data Integration Technologies in Use

Spreadsheets are ubiquitous as a tool for data integration; however, their use can result in versioning, data integrity, quality, and inconsistency issues.

To gain a baseline understanding of organizations' technology choices for data integration technologies, platforms, and services, we asked research participants which ones they are currently using and plan to use in the future (see Figure 3). This report will discuss most of these options further as we move through the research.

Spreadsheets, of course, are common; 68% are currently using them for data integration with 15% planning to do so. Along with functionality for performing calculations, analyzing data, and creating graphs, spreadsheet programs enable users to store and aggregate data. Users typically

share data with others by emailing spreadsheets or sharing work over networks, the web, or collaborative platforms such as Microsoft SharePoint. The freewheeling nature of how users enter data and share spreadsheets can risk versioning problems, data integrity and quality issues, and inconsistent data definitions. Spreadsheet data chaos is a significant pain point for many organizations.

Which types of data integration technologies, platforms, and/or services are currently in use at your organization? Which ones does your organization plan to use in the near future?

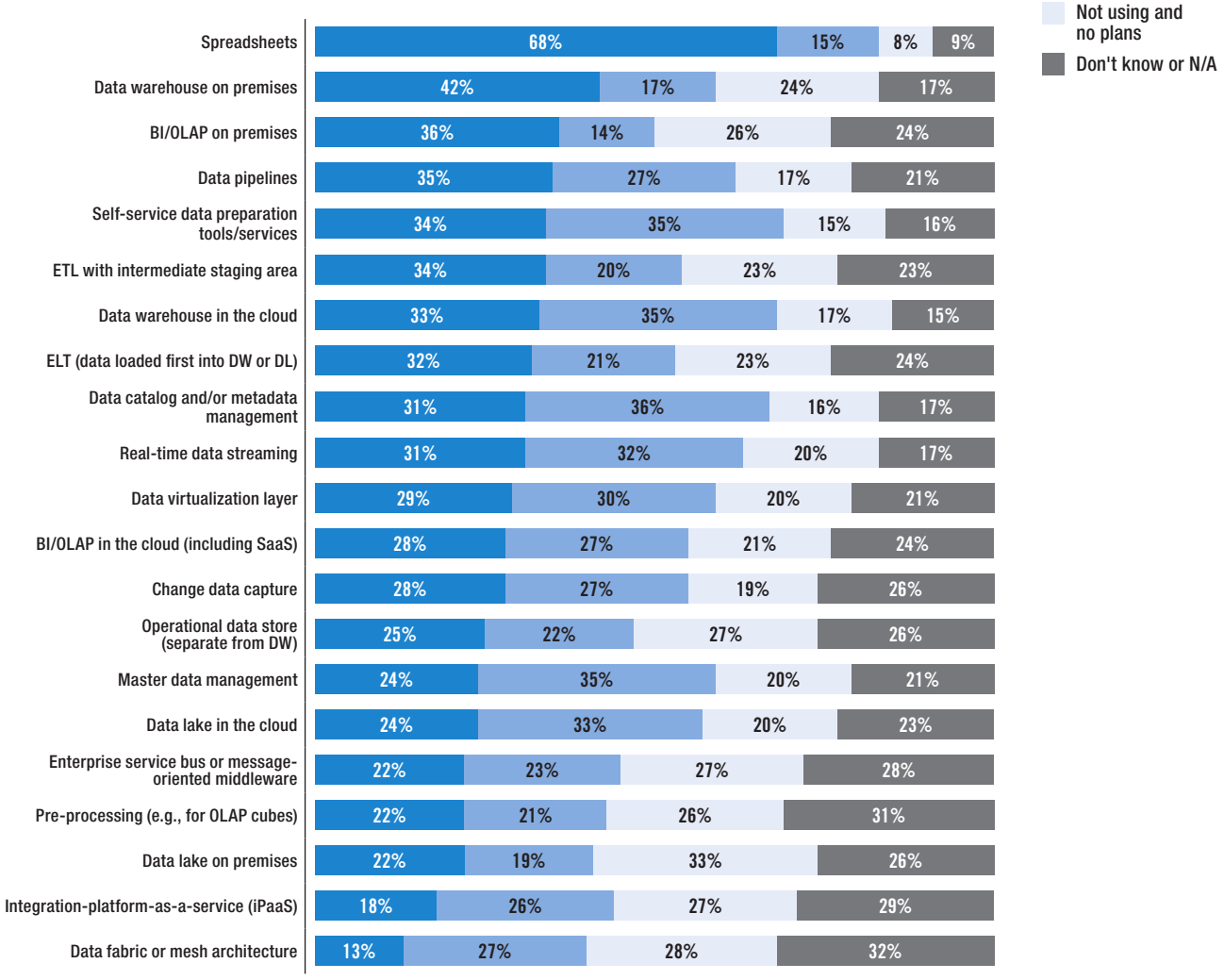


Figure 3. *Based on answers from 415 respondents. Ordered by highest percentage currently using.*

Spreadsheets offer simplicity for aggregating data, but they can become data dumps due to a common lack of formality regarding how users incorporate new data and cleanse, enrich, and prepare it. Spreadsheets often involve considerable manual work, including cutting and pasting between multiple spreadsheets and connecting spreadsheets via custom data pipelines. Novices' inexperience can result in variations between data integration practices. In addition, when users try to integrate new and larger data sources (such as streams of web logs, IoT sensor data, mobile

device data, geolocation data, or data from public web APIs), they typically run into scalability problems in capturing and storing data as well as difficulties trying to analyze it.

Organizations often need to extract data from spreadsheets as part of data integration tasks for reporting and analytics. These efforts can become bogged down with inconsistencies. One key concern is that formulas, calculations, and logic may be lost during data integration. Fortunately, solutions in the marketplace enable users to retain these attributes and apply automation to make extraction and replication from spreadsheets easier and more consistent.

Many research participants are currently using an on-premises data warehouse, but 33% currently have one in the cloud and 35% are planning to have one there in the future.

Data warehouses are on premises and in the cloud. Figure 3 shows that 42% of research participants are currently using an on-premises data warehouse and 33% have a data warehouse in the cloud. Regarding the future, 17% plan to use one on premises but more than double that (35%) are planning to have a cloud-based data warehouse. This shows the attraction of cloud data platforms and growth in the future.

A data warehouse is vital for providing users with properly prepared, structured, and transformed data for BI reporting, dashboards, and many forms of analytics. A standard data warehouse centralizes data from one or more of an organization's business applications, transaction systems, and various external data sources such as third-party credit reporting or customer segmentation services. The goal is for all types of users to have a central, single source of prepared data relevant to subjects of interest.

Having a data warehouse enables users to examine integrated data from multiple business applications and transaction systems without having to go to each source directly. By fielding users' queries, a data warehouse shields source applications having to address them, preserving performance and security. Modern data warehouses optimize queries automatically to improve performance.

Data warehouses offer summarized and aggregated records that can be viewed across a variety of dimensions; users are not limited to the way records are stored in each of the original business applications, processes, and transaction systems. A data warehouse often contains both atomic and summarized data. The data is historical, but modern data warehouses can supply current, near-real-time "live" data.

The consolidated historical data, however, can be extensive and voluminous, particularly in the cloud where systems can take advantage of separately scalable storage and computation processing resources. Because cloud data warehouses can tap as much cloud storage as they need and run on massively parallel processing (MPP) database engines, they have higher scalability than fixed-size, on-premises systems. Organizations can add or remove storage and processing resources as well as adjust access strategies to gain cost elasticity.

Indeed, the modern technology environment in the cloud enables organizations to update data strategies and address problems that have held them back due to the limitations of traditional on-premises data warehouses. Cloud data warehouses primarily support the expected managed enterprise BI reporting and operational dashboards; however, cloud data warehouses have additional flexibility for greater numbers of both users and workloads. Organizations can dynamically manage the volume of "hot" data available for workloads. They can set up zones for complex analytics, such as for developing predictive insights, exploring new data, and creating prescriptive recommendations.

ETL and ELT processes are central to data integration. To consolidate data from multiple sources, most data warehouses depend on extract, transform, and load (ETL) processes or a reordered sequence that loads and then transforms the data (ELT). Just over a third of respondents (34%) say they are currently using ETL and 20% plan to do so. ETL data integration involves extracting data from original sources to a specialized staging area where processes follow business rules to cleanse and transform the data before the next phase, which is loading it into a data warehouse or other target repository.

Data transformation is critical to meeting user and application data requirements. ETL is still common, but 32% are using ELT and 21% plan to use it.

Broadly, transformation is how users and applications get data structured, cleansed, validated, and enriched so it is appropriate for the use case. At this step, processes use rules and data mapping information to cleanse, join, and enrich data from different sources for reports, interactive dashboards, and analytics. Most users and applications need cleansed and transformed data so they can view and access integrated data sets and move faster to apply data to business decisions and collaborate with the data confidently.

ETL programming traditionally requires IT-level expertise to manage numerous processes (possibly hundreds or thousands) of varying complexity. As a result, ETL processes can be a bottleneck and require manual IT work that nontechnical users cannot perform to develop routines and fix problems. However, easier-to-use self-service interfaces and AI augmentation in some data transformation solutions enable users to do more of their own projects without IT intervention. Figure 3 shows that 34% are using self-service data preparation tools and services, which can include ETL processes; 35% are planning to use them.

In Figure 3, nearly the same percentages of respondents are currently using ELT (32%) or plan to use this variation (21%). With this model, organizations build data pipelines to load (or replicate) data into a target repository before transforming it; 35% are currently using data pipelines and 27% plan to use them. Rather than involve a separate staging area, the target repository for ELT is the data platform—the data warehouse itself, an operational data store (ODS), a data lake, or a unified data architecture that combines them. If there are multiple original sources, organizations will often extract and load data first into a data lake, particularly a cloud data lake built on cloud storage.

The ELT option reduces data latency caused by moving or replicating data between a separate staging area and the target platform. ELT can also lower data egress charges incurred when data moves from one network to another, which is important for managing cloud data management costs. ELT has the advantage of being able to use the (increasingly cloud) data platform's powerful and scalable MPP database engines for fast and reliable performance of complex data transformation and cleansing.

The ELT option reduces data latency caused by moving or replicating data. ELT can use powerful database processing engines, particularly in the cloud, for complex data transformation and cleansing.

Data preparation and integration within BI/OLAP platforms draws interest. BI/OLAP platforms typically sit on top of data architectures and make strong use of underlying data warehousing, ETL, or in-database ELT. Some today offer a semantic layer that enables direct access to the data lake. In recent years, BI/OLAP platform solutions have incorporated some data preparation and ETL functionality. Although limited compared to dedicated data preparation solutions, BI/OLAP platforms integrate the functionality they do offer into semantic layers to support self-service analytics model development, data set selection, search, and visualization. To provide scalability and greater functionality, the solutions also provide integration with third-party dedicated solutions for data integration and blending, ETL/ELT, and data catalogs.

Semantic layers are critical to the quality and consistency of business representations of data for reporting, calculations, and multidimensional modeling. Over a third of respondents (36%) say their organizations currently use BI/OLAP on premises for data integration, with 14% planning

to use this mode. As with data warehousing, we see a strong trend toward the cloud, with 28% using BI/OLAP in the cloud (including SaaS) and 27% planning to do so.

Operational data stores to provide views of the latest data. In addition to a data warehouse, some organizations establish an ODS. One-quarter of respondents (25%) currently use an ODS as part of their data integration strategy and 22% plan to use one.

As the name suggests, an ODS provides snapshots of operational data drawn from multiple sources. An ODS is particularly valuable for giving users the most recent view possible of the data. For example, hourly (or even more recent) sales records could be extracted from sources and loaded into an ODS to give managers the latest sales performance. Like a data warehouse, an ODS shields operational and transactional systems from having to field queries directly; users who want to know the latest status will query (or receive results pushed from) the ODS.

Unlike a data warehouse, the "volatile" ODS overwrites previous data as it is updated rather than storing it as historical data records. Thus, ODS data values are constantly changing. Organizations typically update data warehouses using batch processes that insert records into existing tables as part of a growing "nonvolatile" historical data set.

Historically, an ODS became necessary because users wanted access to data snapshots ahead of long transformation, cleansing, and loading processes. Some organizations, however, chose to do light transformation and cleansing of the data before loading it into an ODS to serve carefully limited and repeatable operational reports and dashboards. An ODS can double as an ETL staging area that allows users to explore and profile new raw data sets, which, if found worthwhile, are then transformed and loaded into a data warehouse. Organizations can use an ODS to troubleshoot quality problems in data sources or look for errors in data pipelines and extraction processes.

When not previously transformed, ODS data is similar to that found in a data lake: it is fine-grained, non-aggregated, and not cleansed. Taking advantage of modern storage and processing technologies (particularly in the cloud), some organizations choose to incorporate ODS functionality into their data lake. Data lakes are typically set up primarily to support analytics and AI/ML interaction with vast stores of raw data in its native format. To support ODS-style functionality, organizations can set up zones inside data lakes that consolidate data feeds from selected sources and update this data continuously for operational reporting.

Growth anticipated for data lakes in the cloud. 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 for analytics and AI/ML. Some organizations use a data lake as a staging area for data transformation and cleansing before loading data into a data warehouse. Others use technologies to create operational data lakes that ingest data, including real-time streams from various source systems for operational reporting and real-time analytics.

TDWI anticipates growth in data lakes in the cloud; 24% currently have a data lake in the cloud and 33% plan to have one. A significant percentage are also using or plan to use data virtualization.

In Figure 3, 22% of respondents currently have a data lake on premises and 24% have one in the cloud; one-third of organizations (33%) surveyed plan to have a data lake in the cloud and 19% plan to have one on premises. By basing data lakes on cloud object storage, organizations take advantage of cost-effective, independently scalable storage and processing, which is important given that large data lakes easily reach into the petabytes in data volume.

Historically, data lakes have been the province of data scientists and programmers who know how to work with raw data, including real-time streams, and set up custom data pipelines with embedded cleansing and transformation routines. Originally built using open source Apache

Hadoop ecosystem technologies, data lakes have been essential for analytics-driven NoSQL workloads—for example, those that deliver advanced personalization recommendations based on unstructured data. Such workloads typically demand a variety of customer behavioral data, demographics data, location intelligence, and contextual data. Setting up a data lake in the cloud on object storage offers a faster path to launching data science projects for exploratory, predictive, and real-time analytics.

A significant percentage currently uses or plans to use data virtualization. Figure 3 shows that 29% of respondents' organizations are currently using a data virtualization layer and 30% are planning to use one. 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.

A data virtualization layer enables BI and analytics users, as well as applications and artificial intelligence (AI) programs, to transparently access numerous data silos typically present in hybrid, multicloud environments without having to wait for the data to be moved into a data warehouse or data lake. Modern data virtualization solutions have invested in query optimization to ensure the shortest path to and from the data, minimizing network traffic, a critical matter in making cloud computing fast and cost-effective.

Building on data virtualization is the notion of a data fabric; 13% of respondents are currently using a data fabric and about twice as many (27%) are planning to use one. A data fabric is emerging as an important strategy as organizations put more data in multiple cloud-based storage platforms, which can add to existing on-premises data silos.

In the industry, a data fabric or data mesh is variously called an architecture, a framework, and, of course, a fabric. The strategy is to provide a universal and holistic approach to integrating diverse components of physically distributed data environments. Ideally, a data fabric will use services to integrate the necessary components so that data flows easily and users and applications do not have to use specialized code to access each data silo.

More than one-third plan to use data catalogs and master data management. Figure 3 shows that 31% of respondents' organizations are currently using a data catalog or other form of metadata management and more (36%) are planning to do so. Data catalogs, business glossaries, and metadata repositories collect critical information about how data is defined and modeled and where it is located.

Some data catalogs go beyond basic information to gather knowledge about data lineage: the data's origin and what has happened to it during its life cycle, including how it has been transformed, replicated, and shared. Data catalogs play a key role in data governance and overall data management as well as making it easier for users to search for and find data from diverse sources. Data virtualization layers and data fabrics use metadata extensively. These layers can be integrated with data catalogs, combining to create an independent semantic layer that enables viewing, querying, and governing data located in physically distributed locations.

About one-quarter of respondents' organizations (24%) have master data management (MDM) and 35% are planning to have it. Master data is higher-level data about products, suppliers, customers, and other business entities. Master data tends to be stable, which makes it useful for understanding how other types of less-stable data relate to a particular business entity. MDM enables organizations to establish "golden" master data definitions that are consistent, valid, and

accurate across applications and services. With an MDM system, business users have single views of relevant data without having to know, at a technical level, how to access the data and find their way through lower-level database structures.

Organizations see the business benefits of reducing or eliminating data latency. Support for faster updates, continuous refreshes, and real-time analytics are priorities for many respondents' organizations.

Data integration technologies are embracing real-time data. Many organizations see the business benefits of reducing or eliminating latency between data's generation and sourcing and when people and applications can access the data. Predictive and real-time insights from fast data are increasingly critical in enabling organizations to respond proactively to changing conditions and unexpected events. Accomplishing faster updates and real-time analytics is the focus of a range of technologies that are applicable depending on the use case.

Change data capture (CDC) is one well-established technology providing an alternative to delays caused by batch replication and slow ETL processes. CDC 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 data with minimal disruption to operational and transaction systems. CDC is currently used by 28% of respondents and 27% plan to use CDC.

Nearly a third of respondents (31%) say their organizations use real-time data streaming and 32% plan to use it. Where older solutions offered near-real-time data updates, solutions in this area have progressed to where users and applications can access data within seconds or microseconds of the data's generation or publication.

Data streaming solutions offer continuous flows through pipelines that gather and process data right as sources generate it. Users as well as algorithms can then analyze stream data and take a range of actions, including transformations and enrichment for specific use cases. With real-time data from logs, devices, and sensors streaming into data systems, organizations need modern tools and data platforms to support powerful operational analytics that answer complex business questions.

Figure 3 shows that nearly a quarter of respondent organizations (22%) use an enterprise service bus (ESB) or message-oriented middleware (MOM) technologies, with 23% planning to use them. These venerable technology architectures still play an important role in enabling communication between distributed applications, components, and modules. Newer software development patterns such as microservices, containers, and application programming interfaces (APIs) have shifted attention away from ESBs and MOM systems, but many organizations still need them to support integrated communication between legacy applications and components and between older systems and cloud-based SaaS.

Cloud migration is driving development of iPaaS. Demand for integration between multiple cloud platforms appears to be driving ESB's evolution into integration-platform-as-a-service (iPaaS) solutions. Based in the cloud, iPaaS solutions offer a single, unified platform for enterprise data integration, application integration, API management, and other functionality. Organizations use iPaaS solutions to develop, deploy, manage, and monitor integration processes through a single user interface.

Some iPaaS solutions embed AI techniques such as machine learning and natural language processing to improve scalability, flexibility, and ease of use. Figure 3 shows that 18% are currently using iPaaS solutions and 26% plan to use them.

Data Management Systems in Use

Along with technology choices organizations are making for data integration, our research examined the types of data management systems, platforms, and cloud services that organizations currently use and plan to deploy. Data management systems, essential for getting value from diverse data for many users efficiently and securely, vary considerably. Just when relational database management systems (RDBMSs) emerged over hierarchical, network, and object-oriented DBMS competitors to become dominant and widespread, the big data revolution spurred a new technology wave.

Frustration with RDBMS modeling constraints and scalability limitations for big data as well as licensing costs for new instances led to growth in NoSQL data management such as graph databases, column-oriented databases, document databases, and key-value stores. Much innovation occurred within the open source Apache Hadoop and later Apache Spark ecosystems. Today, cloud computing is a major driver behind data management innovation.

Frustration with RDBMS constraints and scalability limitations opened the way for NoSQL data management, including graph databases, column-oriented databases, and key-value stores.

In our research results, spreadsheets are of course ubiquitous, as they were for data integration; 73% are currently using them for data management and 14% plan to use them (Figure 4). Spreadsheets let users store and analyze data, but in flat, nonrelational tables or worksheets. Users who have complex data that demands multiple tables often find that they must move to an RDBMS, perhaps initially a desktop version such as a simple database that runs on a single PC or similar computer.

Our research finds that 33% currently use a desktop database and 12% plan to use one. However, cloud computing is changing the landscape of desktop databases, enabling users to connect more easily to a larger RDBMS to query data. Some organizations are replacing desktop database systems altogether with data management services hosted in the cloud.

Just over half of respondents (54%) are using an RDBMS and 15% plan to use one. Built on the relational model and relational algebra set out in 1969 by E. F. Codd, RDBMSs along with SQL remain the most established data management technologies. RDBMSs are common for online transaction processing (OLTP). They are also used for some data warehouses and for systems geared to handle mixed workloads featuring BI reporting and analytics as well as OLTP-style data recording, updating, and retrieval. However, many data warehouses use dimensional modeling, relational online analytics processing (ROLAP), and multidimensional OLAP approaches.

A large percentage of respondents (60%) say their organizations currently use an analytics or BI platform for data management (and 19% plan to use one), which often includes OLAP functionality. Specialized platforms for analytics or BI integrate what had previously been disparate processes for collecting, integrating, preparing, modeling, analyzing, and visualizing data. Ease of use is increasingly important as the spectrum of users broadens beyond traditional developers, data analysts, and business analysts to include nontechnical but data-savvy business users and data scientists.

Search and natural language query capabilities are becoming important to bridge skills gaps and enable different types of users to uncover insights without technical skills. Leading platforms also support embedded visualization and analytics inside larger or vertical business applications.

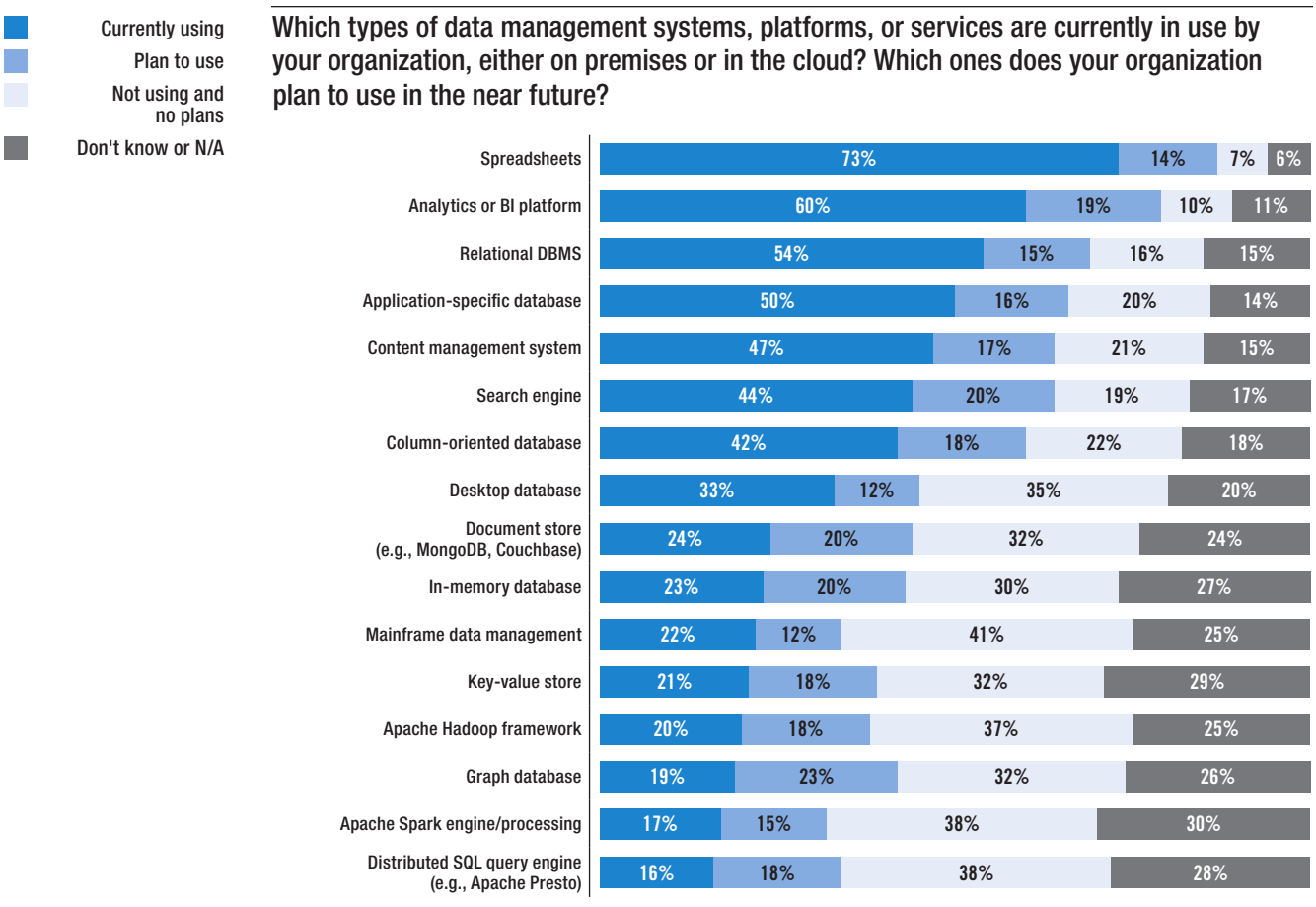


Figure 4. *Based on answers from 403 respondents. Ordered by highest percentage currently using.*

MPP and in-memory capabilities are important. Modern systems located on premises and in the cloud take advantage of massively parallel processing (MPP), often in horizontally scalable, shared-nothing, distributed computing architectures where each processing node is essentially independent. Some data platforms offer both shared-disk and shared-nothing within a single database architecture to suit different types of workload demands.

In-memory databases are also important. These systems keep large amounts of data in main memory; they reduce response time by eliminating the I/O needed to access data stored on disk or solid state drives (SSDs). In-memory DBMSs are good for real-time analytics and operational use cases that require very fast response times.

Compression techniques enable systems to pack more data into memory and make effective use of column-oriented DBMSs. Some in-memory DBMSs protect data from failures by persisting operations on disk in a transaction log. Among respondents, 23% currently use in-memory database management and 20% plan to use such a system; 30% have no plans for it and 27% either don't know or answered "not applicable."

Column-oriented databases are popular. Just over two in five respondents (42%) say their organizations use column-oriented (or columnar) databases and 18% plan to use one (22% have no plans and 18% don't know or answered "not applicable"). Traditional RDBMSs organize and store data physically in a sequence of rows with all columns in each row stored together. This

is good for OLTP and works for managed reporting workloads, but to reach data in numerous columns, complex analytics queries could require sequential scans of what might be enormous tables. This results in poor performance and the retrieval of rows of unneeded data.

A column-oriented database addresses this problem by storing data tables by column. This enables faster access to the data required to answer a complex query and increases efficiency by reducing the amount of data retrieved. 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 DBMSs often use MPP and, with the advent of cloud computing, can take advantage of scalable and cost-effective cloud storage and processing. Although organizations should examine trade-offs depending on their workloads and data warehouse designs, column-oriented DBMSs have proven popular both on premises and in the cloud for large data warehouses supporting complex analytics.

Complex analytical queries benefit from column-oriented databases as well as the use of compression and scalable massively parallel processing.

Distributed SQL engines offer alternatives for querying multiple data sources. Most large organizations have important data distributed across multiple sources. Industry innovations such as open source Presto (originally developed at Facebook and now supported by the Presto Foundation, under the Linux Foundation) are providing distributed SQL engines. Presto and other technologies are similar to RDBMSs and run on clustered MPP systems.

Presto was developed to handle a high volume of analytics queries that interact with massive, multipetabyte data warehouses and other data stores. In a single SQL query, the engine can retrieve and interact with data at multiple sources ranging from other relational DBMSs to proprietary data stores. In our research, we can see that distributed SQL engines such as Presto are not yet in widespread use, with 16% of respondents saying their organizations are currently using one and 18% planning to do so (38% are not using one and have no plans for it and 28% don't know or find it not applicable).

NoSQL Data Management: Current Use and Plans

NoSQL (or not only SQL) is a term that groups a variety of non-tabular data management systems that do not follow the relational model. Although some have been available for a long time, many became prominent as organizations sought to manage larger and more varied big data volumes. Organizations found RDBMSs too inflexible for big data and for use cases with less stable, frequently changing requirements. (Note that some in the industry categorize column-oriented DBMSs as NoSQL.)

The NoSQL trend took advantage of decreases in the cost of storage in the cloud, which had been a longstanding DBMS development constraint. The advent of lower-cost storage arrived at the same time that organizations became interested in unlocking value from massive data that was semistructured, unstructured, and polymorphic. Cloud computing has delivered flexible and low-cost object storage and enabled faster NoSQL processing on horizontally scalable MPP clusters.

Key-value and document NoSQL databases (or "stores") are popular. Using a simple data model of keys and values, key-value databases let users choose keys and query values stored in the database. Just over one-fifth of respondents (21%) say their organizations are using a key-value store database and 18% plan to use one (32% have no plans and 29% don't know or say it is not applicable).

Research participants are using NoSQL; 21% have a key-value store database, 24% have a document database, and 19% use a graph database (23% plan to use one).

Document databases store values in documents, which could include a variety of items such as blogs, social media, and online customer comments. Sets of documents form collections, and one or more collections are contained in top-level folders. Document databases offer transparent, internal data structures; schemas can be represented as JavaScript Object Notation (JSON)

objects and use common JSON or the Apache data serialization system Avro for efficient storage and data exchange. Users can query entire pages of data or look for specific items.

Document databases offer flexibility in how users group documents into collections or domains; queries can retrieve multiple document database entries in a single request. About one-quarter of respondents (24%) are currently using a document database and 20% plan to use one (32% have no plans and 24% don't know or say not applicable). Almost half of respondents (47%) say they use a more traditional content management system (CMS) and 17% plan to use one.

Knowledge graphs and graph databases help organizations store, explore, and analyze data relationships at scale.

Organizations use graph databases to uncover and explore data relationships more fully. To search for and analyze data relationships effectively, organizations need models and systems that can store data relationships and make them discoverable. Knowledge graphs, network-based representations, and graph databases specialize in these capabilities. Knowledge graphs, which are built using graph databases and ontologies, capture how different data sets relate to each other and to higher-level entities such as people, places, and things. The use of knowledge graphs, graph models, and network-based representations is not new; search engine providers use knowledge graphs to increase the accuracy, speed, and completeness of search results.

Graph databases and query languages manage and retrieve data relationships. A graph database, a type of NoSQL database, is an alternative to an RDBMS for storing complex data and exploring and analyzing data relationships across large numbers of data points. Using a graph database, users do not have to modify graph data models to fit a relational database's normalized table structure.

Graph databases enable developers to avoid having to write complex SQL JOIN statements to discover associations in relational databases or program special routines to convert graph structures into relational structures. In our research, 19% of respondents currently use graph databases and 23% plan to use one; 32% have no plans for graph databases and 26% don't know or find them not applicable.

One-fifth of respondents manage data with Apache Hadoop. Finally, we find that 20% of respondents' organizations currently use Hadoop framework technologies for storing and managing data on commodity hardware clusters; 18% plan to use them and 37% have no plans for them (25% don't know or indicate not applicable). Hadoop framework systems exploded on the scene with the growth of web-generated data and the need for scalable, distributed, low-cost data management that could store and swiftly process high-volume data for new types of workloads, such as search engines.

Hadoop-based data lakes have enabled data scientists to explore a variety of new data types and develop predictive models that run on massive data. Along with the Hadoop Distributed File System (HDFS) itself, technologies important for Hadoop ecosystem data management have included the parallel processing framework MapReduce, HBase for data storage and real-time processing, the Hive query engine for data warehousing, SQL-like Pig for query and analysis of complex data structures, and Sqoop for ETL processes.

Over time, development of new open source technologies and frameworks plus cloud computing have changed the landscape, superseding Hadoop and turning some technologies such as MapReduce into legacies. Apache Spark framework technologies, libraries, and modules, for example, enable organizations to unify formerly disparate SQL interaction, near-real-time data streaming, graph processing, and AI/ML development. Spark offers higher performance than Hadoop on massive, fast, and complex data. However, as a newer framework, Spark use is not widespread. Just under one-fifth of respondents (17%) are currently using Spark framework technologies and 15% plan to use them; 38% have no plans for them and 30% don't know or say they are not applicable.

Data Integration: Experiences and Challenges

With a view of the data integration and management systems and cloud services organizations have in place and plans for the future, we can turn our attention to experiences and satisfaction. Then, we'll examine major challenges in meeting objectives such as reducing latency and improving efficiency.

Nearly one-quarter of respondents (23%) say they are very satisfied with traditional ETL and over a third (36%) are somewhat satisfied, with ETL processes and technologies needing some improvement (see Figure 5).

Satisfaction is highest for ETL, although many research participants say they need improvements. Respondents are also relatively satisfied with data collection, integration, and blending.

How satisfied is your organization with its current data integration and management for addressing the following requirements? Which areas need improvement?

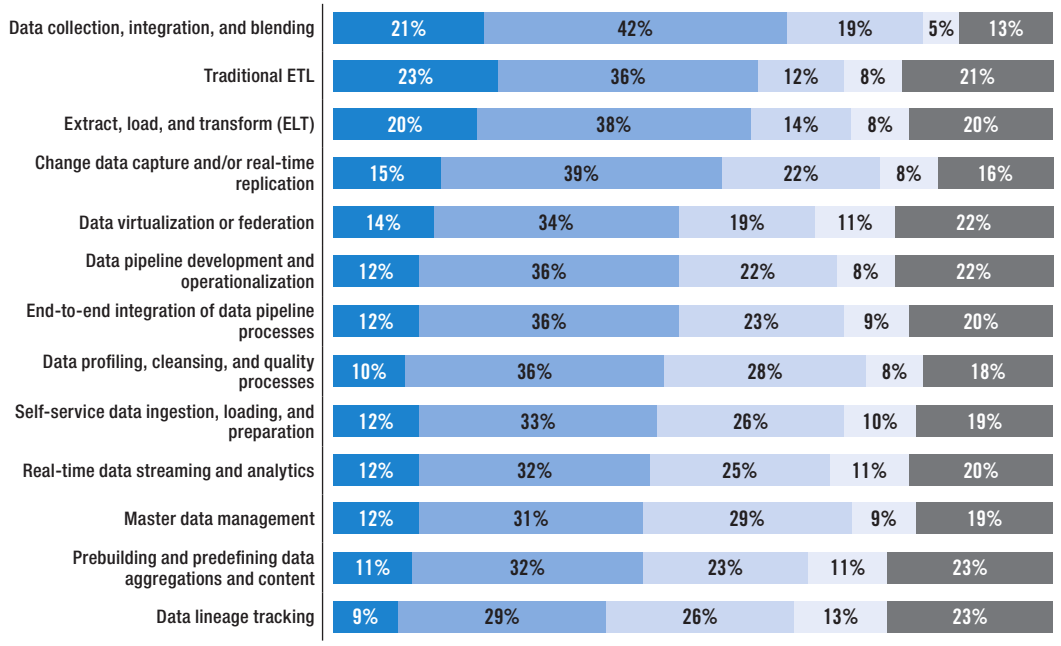


Figure 5. *Based on answers from 373 respondents. Ordered by highest combined "very satisfied" and "somewhat satisfied; needs some improvement" responses.*

Research interviews for this report find that three common challenges with traditional ETL processes drive organizations to modernize:

- **Inflexibility.** ETL processes can be brittle and difficult to alter once data or requirements change, such as users wanting to add new data columns or dimensions. Users frequently want to view or access a wider variety of data sources and types.
- **Lack of automation.** Traditional ETL requires manual work, causing projects to run over budget and take too long. If failures occur during ETL processes, IT experts typically must restart them, bringing more delays. Smart automation in modern solutions can remedy these issues.
- **Data trust, quality, and consistency**. Users are not sure if they can trust the results. ELT processes can lack visibility to check for errors.

As mentioned previously, changing the sequence to ELT is an alternative, particularly when scalable cloud data platforms are available. One-fifth of respondents (20%) are very satisfied with ELT and 38% are somewhat satisfied and need some improvement in their technologies and processes; 14% are looking for a major upgrade.

Ultimately, organizations need to align ETL and ELT approaches with the appropriate use cases. ELT reduces data latency and scalability challenges brought about by moving data to a separate staging area for transformation and then to a target data warehouse. ELT can lower data egress charges incurred when data moves from one network or cloud platform to another. Yet, ETL can be a better choice for some use cases because it typically limits extraction and loading to just the relevant data, for example, for regular, preplanned reports and dashboards. ELT is a better match for deeper, more wide-ranging analytics.

Respondents indicate moderate satisfaction with data collection, integration, and blending. Data collection, integration, and blending are related and somewhat overlapping terms. *Data integration* is the broadest, covering a range of processes for producing valuable, integrated data sets for user views or access. Traditional data collection processes are targeted and intentional; users know what questions they are pursuing and the quality of the data they need. With data collection, variables for analytics are consistent, well-defined, and aimed at formulating credible answers to business questions.

Data blending describes a shorter path compared to data collection and long ETL processes: a faster merging of data from multiple and sometimes diverse sources. Rather than creating a new persistent data warehouse or data mart, blending aims at making integrated data accessible at the visual analytics or reporting layer. Data blending solutions have matured to offer broader capabilities to meet a wider variety of use cases. However, data blending focuses primarily on the needs of self-reliant, data-savvy "citizen" data scientists who want a faster and more flexible path to viewing integrated internal and external data. Users sometimes employ data blending as part of fast prototyping to determine if data sources have enough repeatable value to apply longer, traditional ETL processes for inclusion in a data warehouse.

In Figure 5, data collection, integration, and blending received the second-highest "very satisfied" response at 21%, but twice as many (42%) are only somewhat satisfied and feel these areas need some improvement. Nearly one-fifth (19%) want to see a major upgrade.

Data pipelines need improvement. A data pipeline is a system that enables data to get from its original sources to users' data platforms for BI reporting, dashboards, analytics, and AI/ML development. Over one-third of respondents (36%) say they need some improvement with their data pipeline development and operationalization and 22% are looking for a major upgrade. To improve data pipelines, users may need better training and the ability to share best practices and knowledge about data sets. They also most likely need improvements, if not major upgrades, to their tools and technology platforms to handle scale and speed requirements.

Many data pipelines focus primarily on moving data as fast as possible, including through real-time data streaming to data lakes or other storage. Such pipelines primarily serve data science exploration, AI/ML, and deep learning techniques. However, when use cases demand more than just raw, real-time data streams, pipeline development teams will incorporate data quality and ETL steps; invoking these steps could occur during data acquisition and collection or later. Pipelines can also incorporate governance, user authentication, and security rules and constraints.

When it comes to end-to-end integration of data pipeline processes, the same percentage of respondents (36%) need some improvement and 23% are seeking major upgrades to afford better integration. As pipelines tap more data sources and increase in complexity with the addition of data cleansing, transformation, enrichment, and governance steps, they often involve a variety of point solutions and user processes. The lack of integration between tools and processes becomes a significant source of latency.

Thirty-six percent are looking for better end-to-end integration of often disconnected data pipeline processes and 23% say they need a major upgrade.

The research indicates that many organizations struggle to monitor and manage data pipelines in an end-to-end fashion. Remedies could include instituting frameworks such as DataOps (to improve pipeline orchestration and collaboration between developers) and upgrades to modern, more unified solutions that feature end-to-end integration.

In Figure 5, 28% are interested in a major upgrade to the data profiling, cleansing, and quality processes themselves; 46% show satisfaction with them, but 36% want at least some improvement. Scalability, speed, agility, and completeness are typically the biggest challenges regarding these processes. Organizations need to take advantage of modern solutions that offer increased automation and embedded AI to reduce manual work as data volumes rise and user requirements vary.

The majority spend more time on data integration than analytics. In a related question, we asked respondents what percentage of total project time was spent on data integration, pipeline development, and preparation compared to the time users spent on analytics and visual data interaction. Just as our research in previous years found, the results for this report indicate that users spend more time on these tasks than on analytics activities closer to realizing business value from the data.

Thinking of your organization's most recent BI and analytics projects, what percentage of the total project time was spent on data integration, pipeline development, and preparation?

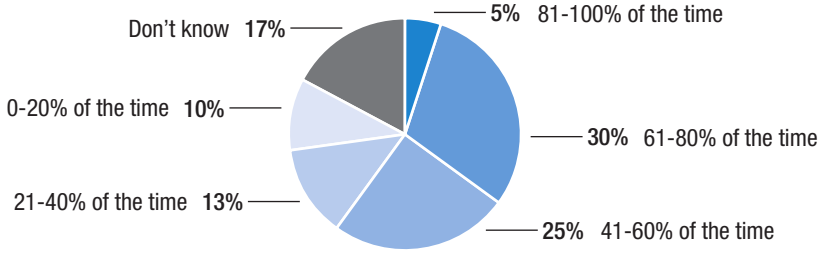


Figure 6. *Based on answers from 375 respondents.*

Technology solutions that apply AI-infused automation are key to enabling organizations to reduce the percentage of time spent on data integration, pipeline development, and preparation. Modernized technologies and updated practices can reduce costly IT burdens and enable users to do more on their own.

About half are satisfied with data virtualization. Some organizations use data virtualization and federation as either a complement to or as an alternative to data pipelines and the heavy data movement involved in ETL processes for data warehousing. These organizations use a data virtualization layer to connect to and query multiple and distributed sources, aggregate the resulting data, and provide views of it from a single point of access.

About half of respondents (48%) are satisfied with data virtualization, which enables users to connect to and query multiple and distributed data sources from a single point of access.

Just under half of respondents (48% in Figure 5) say their organizations are at least somewhat satisfied with data virtualization, with 34% saying they need some improvement; 19% express interest in modernizing through a major upgrade. Because data virtualization differs from ETL

and data warehousing, organizations often need training in how to apply virtualization to their workloads effectively. Organizations' administrators may need upgraded tools and training so they can manage virtual images and ensure good performance and security as more are created.

Organizations are interested in modernizing master data management. MDM helps organizations understand how data is relevant to higher-level business entities and hierarchies. Master data definitions ideally are consistent, valid, and accurate across multiple databases, applications, and services. MDM is often a challenge to establish, but it can improve data integration and collection. Figure 5 shows that only 12% of respondents are very satisfied with their current MDM, reflecting the difficulties of defining master data and integrating underlying definitions; 31% are somewhat satisfied, needing some improvement, and 29% are seeking a major upgrade to their current skills and available MDM solutions.

A major MDM modernization focus is to move beyond single-domain MDM to multidomain MDM. Organizations have historically invested in single-domain MDM systems serving specific data management systems or business applications such as ERP, CRM, customer information management (CIM), supply chain management, and financial management. Multidomain MDM centralizes single domains into one consolidated MDM system or cloud service. This increases management efficiency and reduces confusion. Organizations can gain single views of complex, higher-level business entities across domains for concerns such as pricing, supply chain resilience, and product development.

Steps for Improving Data Pipelines and Preparation

As noted earlier, end-to-end integration of data pipeline and data preparation processes is a major challenge. Thus, it is not surprising in Figure 7 that the highest percentage of respondents (45%) regard end-to-end integration of data ingestion, cleansing, transformation, validation, and enrichment as an important step they could take to improve data pipelines and data preparation. Nearly a third (31%) want to improve the speed of data pipeline and preparation steps, and one-quarter of respondents (25%) say they need to orchestrate development and execution of growing numbers of data pipelines. Nearly the same percentage (23%) want to reduce the cost of data pipelines and data preparation.

The results suggest that organizations would benefit from implementing DataOps, although only 11% indicate that it is important to implement such a framework. DataOps borrows from agile and DevOps methodologies to give organizations a framework for eliminating delays and inefficiencies in disconnected phases of data life cycles, including development, data collection and integration, quality, profiling, transformation, enrichment, governance, and capacity planning. DataOps has proven to help organizations orchestrate data pipelines, contain costs, and apply modern technologies for improving end-to-end integration.

Which of the following steps are most important to your organization for improving data pipelines and data preparation? Select up to five.

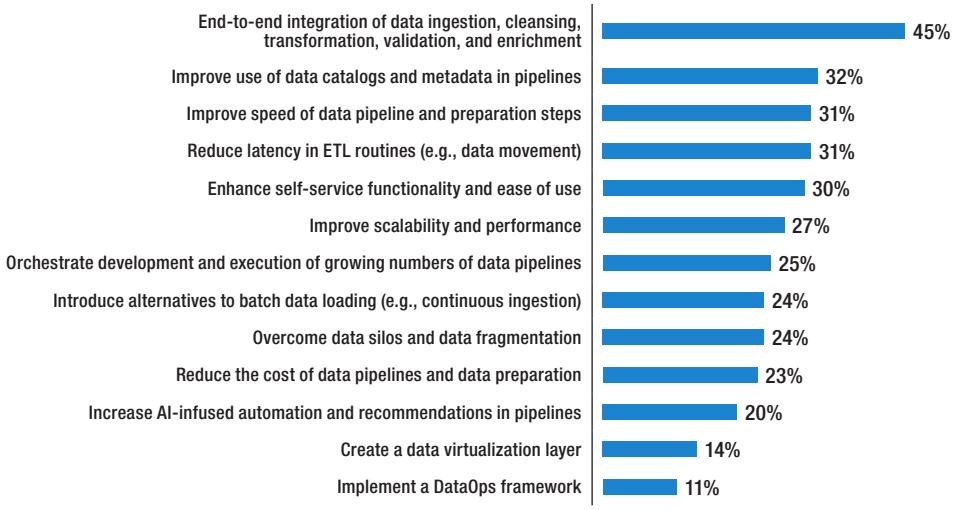


Figure 7. *Based on answers from 364 respondents.*

Many organizations prioritize reducing latency in data movement and ingestion. Figure 7 shows that 31% of respondents' organizations regard reducing latency in ETL routines to be a priority for improving data pipelines and preparation. ETL latency challenges typically increase as data volume and variety rise and more users are demanding fresh data. Data virtualization can be an alternative for providing views of data without data movement; 14% want to improve by creating a data virtualization layer.

Nearly one-quarter of respondents (24%) say that introducing alternatives to batch data loading such as continuous ingestion would be important. Data ingestion can take many forms, including as part of ETL or ELT; the purpose is to get data from sources into target systems in the right structure, format, and quality needed for designated use cases.

To reduce latency, some organizations implement solutions that offer continuous, mass, or streaming ingestion services, often through pipelines. These options take snapshots of data or stream data from CDC, logs, or other sources as soon as it is available. Then, in microbatches, they move data to targets such as messaging systems (e.g., Apache Kafka), cloud data storage, data lakes, or data warehouses.

Nearly one-third want to improve use of data catalogs and metadata in pipelines. Allowing data pipeline processes to access a data catalog or other metadata management in an automated fashion enables users to find data faster, whether the data is located in a physically centralized data lake or in a table managed in a distributed data platform. Leading data catalog solutions keep track of changes to data schema or structure—changes that if not detected can lead to data pipeline failures. Catalogs enable users and administrators to discover data quality problems before pipelines move or replicate data downstream to target destinations. In Figure 7, 32% of respondents regard improvement in the use of data catalogs and metadata in pipelines as a key step.

The amount of hands-on work required by data administrators and IT developers has long prevented data catalogs from playing a bigger role in data pipeline and preparation processes.

In the next section we discuss how modern solutions address challenges in keeping metadata accurate, up to date, and available for data pipelines and preparation.

Data Catalogs and Metadata Management

Diverse users and workloads are demanding access to varied and voluminous data. This increases the need for accurate, well-managed metadata. In this section, we look more closely at how organizations are using data catalogs (if at all) for metadata management. We will explore satisfaction with them for addressing various requirements. We will also discuss organizations' modernization priorities.

A centralized metadata repository such as a data catalog helps organizations reduce confusion about the data and enhances curation and governance.

A centralized metadata repository such as a data catalog helps organizations reduce confusion about the data's quality and consistency and enhances data curation and governance. A complete and accurate data catalog includes what the data means and represents, the data's origins and where to find it, who is responsible for the data, and information about whether it is relevant to users' objectives. Users are able to see relationships between data sets.

Regarding databases, a data catalog documents existing relational schemas, data elements such as tables and columns, and other database objects such as stored procedures. Some data catalog solutions document metadata structures for NoSQL databases, including the tree structure of folders, collections, and schemas in document databases. Solutions let users apply APIs to upload NoSQL database metadata details for display and maintenance from within the data catalog. Organizations that have data lakes also need data catalogs to track data ingestion so the lake does not become an unusable data swamp. Data catalogs can track what happens to the data lake's assets, including transformations for data warehouses or BI and analytics platforms.

Developing an enterprise data catalog as a single source is an important goal, especially as data spreads across a hybrid of on-premises and multiple cloud-based systems. Enterprise data catalogs help organizations overcome data fragmentation problems such as poor data quality and too little governance—problems often caused by having disparate data silos. However, organizations more commonly have catalogs deployed at a lower level, where they still play useful roles in collecting metadata and providing a shared and growing knowledge base. Thus, the role of each data catalog can vary.

TDWI asked organizations about the scope of their data catalog(s) for metadata management.

Does your organization manage metadata in a data catalog (or multiple data catalogs)? If so, what is its scope?

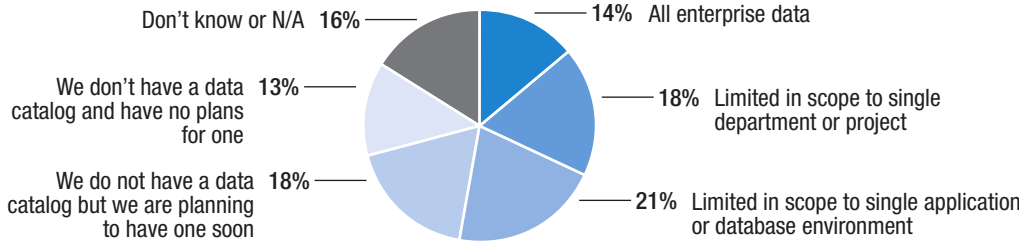


***Figure 8**. Based on answers from 382 respondents.*

Scattered metadata in numerous applications and databases often drives organizations to establish their first data catalog so users, developers, and administrators do not have to move from source to source to discover and use metadata. Metadata across silos is typically inconsistent and haphazard when organizations do not manage it efficiently. Smaller-scale data catalogs can themselves become silos, ultimately necessitating improvement through centralization into an enterprise catalog. Fortunately, modern data catalog solutions can speed up this evolution by crawling disparate data sources and using AI to automatically parse, deduce, and tag credible metadata.

Data Catalog Satisfaction for Important Objectives

A data catalog can offer accessible and up-to-date knowledge about the data, its lineage, and its location. It shortens the time it takes for people to find relevant data and share effective preparation and enrichment processes. An accurate data catalog reduces confusion about data quality and speeds remediation as users connect to new sources.

We see in our research that catalog use for improving data quality, consistency, and completeness still requires modernization; 45% are satisfied, which is the highest satisfaction percentage in Figure 9, but only 11% are very satisfied and nearly one-third are unsatisfied (31%). Organizations increase business benefits when they use data catalogs to coordinate data meaning across sources; 42% are satisfied with their data catalog for this purpose and 12% are very satisfied.

TDWI research shows that 45% are satisfied with data catalogs for improving quality, consistency, and completeness.

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

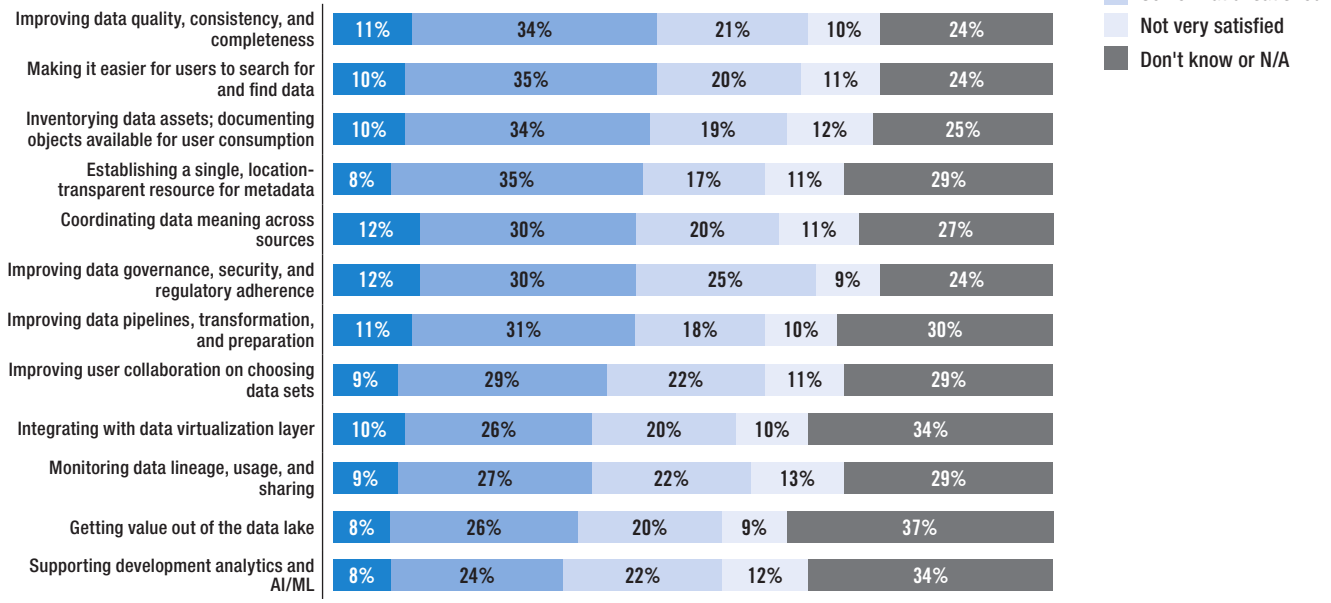


Figure 9. *Based on answers from 379 respondents. Ordered by highest combined "very satisfied" and "somewhat satisfied" responses.*

Nearly half of those surveyed say catalogs make it easier to search for and find data. Figure 9 shows that 45% of respondents' organizations are satisfied with the role of their data catalog in making it easier for users to search for and find data, with 10% very satisfied. At the

point of data use, users can access the data catalog to learn information about the data, including its consistency, age, and governance constraints. Nearly a third (31%) see room for improvement.

For users, improvement could come from modern data catalogs' use of natural language search capabilities to make it easier to find data. These capabilities fit with trends toward supplying AI-driven recommendations about data sets within BI and analytics applications. Some solutions provide data catalog search from within these applications rather than requiring users to access the data catalog directly.

A centralized data catalog makes it easier for physically remote users to collaborate. In Figure 9, we can see that 43% are satisfied with their data catalog for establishing a single, location-transparent resource for metadata, although there is room for improvement as only 8% are very satisfied and 28% are unsatisfied. As noted, organizations can apply features in modern solutions that crawl distributed data platforms for metadata to coordinate data definitions and meaning across platforms.

Some data catalogs automate identification of experts who can act as stewards to field questions about and guide interaction with particular data sets. Searchable, crowdsourced knowledge collected by data catalogs can support informed collaboration among remote users. Rating features in data catalog solutions enable users to share insights about data sets. Figure 9 shows that 38% are satisfied with their data catalog for improving user collaboration on choosing data sets; 9% are very satisfied; 33% see room for improvement, suggesting that these organizations' data catalogs do not have robust collaboration features.

Forty-four percent of respondents are satisfied with their data catalog for inventorying data assets and documenting objects available for user consumption. These capabilities are critical for data governance and data privacy regulatory adherence.

Organizations use data catalogs to inventory data and support governance. As organizations source varied and voluminous data, they need to ensure they have up-to-date inventories of sensitive data—such as personally identifiable information (PII) about customers and consumers—and can track its use. Inventories of other mission-critical data are important to locating data sets faster and overall data management. In Figure 9, we see that 44% of respondents are satisfied with their data catalog for inventorying data assets and documenting objects available for user consumption, with 10% very satisfied.

Data governance and regulatory adherence are high priorities and a key area where data catalogs contribute. Along with an inventory of various forms of metadata about data assets (e.g., structural, technical, business, and usage metadata), data catalogs often contain governance and security rules that drive constraints in data pipelines and applications. Figure 9 shows that 42% are satisfied with their data catalog for improving governance, security, and regulatory adherence with 12% very satisfied. About a third (34%), however, are unsatisfied.

Dissatisfaction is often due to manual work involved in populating data catalogs, classifying data for governance and security, and tracking data lineage—that is, determining data origins and what has happened to it during its life cycle. As part of data lineage, catalogs (as well as specialized data lineage tools) can document how users have transformed, replicated, and shared the data. Just over a third of respondents (36%) are satisfied with their data catalogs for monitoring data lineage, usage, and sharing but only 9% are very satisfied and 35% indicate dissatisfaction.

Modern tools have improved automation for discovering and documenting data lineage and suggesting missing data lineage for data elements that might have come from a BI report, a spreadsheet, an ETL process, or data warehouse, each of which is likely to have its own metadata conventions. Modern data lineage tools and data catalogs provide visual representations that enable administrators and users to see data lineage, spot where information is missing, and track data's flow through the organization.

Modern data catalogs have improved automation for discovering and documenting data lineage. AI-driven automation can embed governance constraints into applications.

AI-driven automation can support the notion of governance "policy as code," i.e., governance constraints embedded in applications so that users receive guidance as they interact with data. Such capabilities enable organizations to have visibility into data use no matter where the data physically resides, such as on premises or in cloud-based data platforms. Location-transparent data catalogs are important for managing and accessing data that cannot be moved or replicated due to a country or region's data residency regulations.

Embedded governance constraints reduce pressure on IT personnel to monitor all activity to ensure that users follow governance rules and the organization complies with regulations. Crowdsourced insights about data sets collected in data catalogs additionally help users collaborate to ensure governance and data quality. Documentation of crowdsourced notes from subject matter experts fills in gaps in data lineage with contextual information and can be used to improve mapping of metadata to business entities.

Data catalogs help data pipelines, transformation, and preparation. In Figure 9, about two in five respondents (42%) are satisfied with data catalogs for improving data pipelines, transformation, and preparation, with 11% very satisfied and 28% unsatisfied. Closer integration of data catalogs with these processes saves users' time in locating relevant data. Automation in data catalogs enables faster identification of data quality and consistency issues as data sets get bigger and more numerous.

AI capabilities help organizations move beyond just keeping data inventories toward establishing data curation processes that increase the data's value and govern and protect it as an asset throughout its life cycle from creation and sourcing through preparation and use. During data pipeline development and transformation and preparation processes, some catalogs can automatically expose which data sets are available, trusted, and governed. These "active" data catalogs depend on AI-driven automation to eliminate delays that occur when passive data catalogs require users to "wake them up."

Data Catalog Modernization Priorities

The metadata management and access provided by a data catalog can be a cornerstone of modern data integration and the foundation for higher-level semantic integration vital to data-driven business objectives. However, a data catalog requires continuous attention and modernization. TDWI asked respondents which modernization steps they regard as most necessary. Here are some key findings:

Automation is a high priority. The highest percentage of respondents (45%) want to increase automation to reduce manual maintenance and development (no figures shown for this topic). As noted, traditional data catalogs have required significant manual work, which has caused them to fall short of delivering full value. Modern solutions automate data catalog development and maintenance.

Respondents also want to see more automation in the discovery and documentation of data lineage (38%). About a quarter of respondents (24%) regard automated scanning and tagging of new data sets for governance as a priority.

Consolidation into a central resource is important. Modernization should address the problem of fragmented knowledge about disparate data sets, including NoSQL data. Nearly a quarter of respondents (24%) want to use their data catalog for semi- and unstructured data as well as structured data. Just over one-third of respondents (34%) say it is important to consolidate data profiles, business rules, quality metrics, and other metadata existing in disparate data systems and applications on premises and in the cloud.

In addition, 27% prioritize consolidation of smaller departmental or project-focused data catalogs into an enterprise catalog. Nearly the same percentage (26%) indicate the importance of easing integration of external third-party data catalogs and metadata management. As users reach out to external data sources through data marketplaces and exchanges, they need APIs to link and align knowledge about different data sets and sources.

One-quarter (25%) of research participants say that integrating data catalog metadata resources with their data virtualization layer is a modernization objective.

Data catalog integration with data virtualization is a priority. One-quarter of respondents (25%) say that one of their modernization objectives is to integrate data catalog metadata with data virtualization layers to improve query speed and breadth. This would enable users to browse the data catalog to locate important data, which could be physically in a centralized data lake or distributed data platforms. Using the data virtualization layer, they can preview or query the data. In Figure 9, we saw that 36% are currently satisfied with their data catalog integration with data virtualization; 10% are very satisfied.

Modernization should enhance user interaction with the data catalog. By using annotations and rating systems to crowdsource and share experiences with data sets, users can take advantage of data catalogs for collaboration. About one-fifth of respondents (21%, not shown) want their data catalog to provide functionality for users to augment metadata, such as to certify and rate data sets; 13% want catalogs to embed natural language functionality to help users ask questions.

Expanding the knowledge base and using AI/ML effectively is important. Some respondents (17%) want to augment their data catalog with AI/ML to increase scalability, accuracy, and speed to support complex queries running across different data sources and types. With use cases becoming more complex and data sources diversifying, some organizations are expanding into semantic data integration. About one-fifth of respondents (19%) want to modernize by incorporating descriptive, semantic knowledge about diverse and distributed data into a more complete knowledge base than just a metadata repository, including by using methods such as knowledge graphs.

Analytics: Driving Integration Innovation

Organizations are excited about the potential of data visualization, analytics, and AI. Analytics embodies a quest to unleash the data—to "let the data speak" so people and applications can operate with fuller, more factual understanding. Many organizations use analytics as an umbrella term that includes BI, OLAP, data visualization, search, predictive modeling, and AI techniques such as machine learning and natural language processing.

Monolithic, one-size-fits-all enterprise data integration and data management platforms often prove inflexible for analytical data interaction.

Moving beyond standard reports and dashboards for simple data consumption, data-savvy business users—often called citizen data scientists—want to explore multiple new data sources from different perspectives. They want single views or access to all relevant data so they can determine causal factors and correlations and interpret changing conditions to achieve best outcomes.

Monolithic, one-size-fits-all enterprise data integration and data management platforms often prove inflexible for more advanced and widespread analytics. TDWI asked research participants about their organizations' satisfaction with current data integration technologies, platforms, and practices for supporting the following activities important to data visualization, analytics, and AI/ML (see Figure 10).

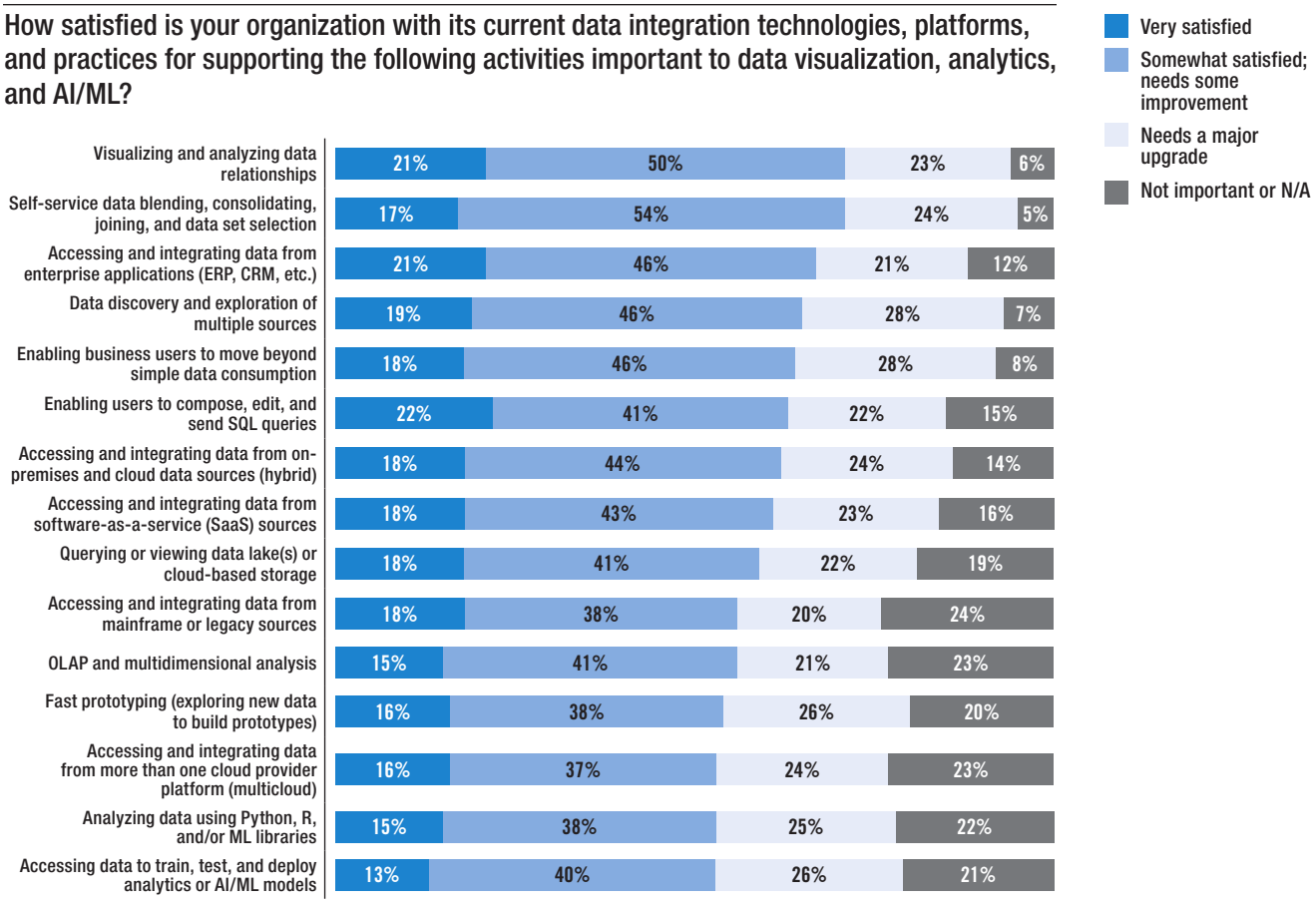


Figure 10. *Based on answers from 424 respondents. Ordered by highest combined "very satisfied" and "somewhat satisfied; needs some improvement" responses.*

The highest percentage showed satisfaction with being able to visualize and analyze data relationships; 21% are very satisfied and 50% are somewhat satisfied, seeing the need for some improvement. This shows the power of current data visualization technologies for enabling users to see relationships in data sets by selecting from libraries of visualization types. Nearly one-quarter (23%), however, say they need a major upgrade, which indicates users need tools that make it easier to integrate and blend data for visual analysis. Some modern solutions use AI to either recommend visualization types or automatically present data in the appropriate visualization.

Visual analytics and data discovery technologies empower users to move beyond simple data consumption. This typically heightens demand for data availability to handle new, ad hoc and complex queries. Satisfaction among respondents is moderate; 18% of respondents say their organizations are very satisfied with current data integration for moving beyond simple data consumption.

The bulk of respondents (46%), however, are looking for some improvement, and more than a quarter (28%) believe they need a major upgrade. Figure 10 shows that respondents have nearly the same levels of satisfaction regarding data integration for data discovery and exploration of multiple sources and for enabling users to compose, edit, and send SQL queries. Stewardship and data literacy training can help users build confidence to move beyond simple data consumption

to do data discovery and exploration. Organizations may need to upgrade data integration and management to support expansion in self-service data discovery and exploration.

Satisfaction with how integration supports OLAP and multidimensional analysis is a bit weaker; just 15% are very satisfied and 41% are somewhat satisfied. These types of analysis demand more expertise in data modeling, data preparation (such as cleansing and formatting), and accessing the data warehouse or ODS to set up data marts and OLAP cubes for slicing, dicing, and rolling up data. Modern BI/OLAP data platforms automate steps in the development and maintenance of OLAP cubes; cloud services are addressing scalability and delays experienced with traditional, on-premises OLAP data integration processes.

Respondents are relatively satisfied with self-service data preparation, but significant numbers desire improvement and major upgrades.

Self-service data preparation satisfaction is relatively strong. Data preparation includes procedures from the initial ingestion of raw data to loading and collecting it into a target platform and transforming and enriching the data into a usable state for analytics, AI/ML, BI, and other applications. Self-service data preparation is an important trend for enabling data scientists, analysts, developers, and data-savvy business users to do more on their own.

In Figure 10, 17% of respondents say their organizations are very satisfied with how their data integration practices support self-service data blending, consolidating, joining, and data set selection, and 54% are somewhat satisfied but want improvement. Nearly one-quarter (24%) need a major upgrade. Technology advances can provide improvement and upgrades through automation of data preparation processes in pipelines to reduce manual work in data selection, profiling, cleansing, transformation, and validation.

Organizations need improvement to support advanced analytics. Most respondents want better data integration for data science. Just over half (53%) are satisfied with data integration to support use of Python, R, and/or ML libraries; 38% are only somewhat satisfied and need some improvement, and 25% indicate a major upgrade is required. This suggests that users need training in libraries, tools, and packages for data integration, manipulation, cleansing, and enrichment when they are using Python, R, and ML libraries. Modern, integrated data science toolsets make it easier to undertake data collection tasks for training predictive models and building ML algorithms.

Unified data architectures that integrate data lakes and data warehouses into a data lakehouse could provide improvements for supporting a wider range of programs beyond standard SQL as well as access to a wider variety of data. Figure 10 shows that just under one in five respondents (18%) are very satisfied with users' ability to query or view the contents of data lake(s) and cloud-based storage; 41% are somewhat satisfied and 22% need a major upgrade.

As sources proliferate, data access and integration satisfaction varies. Respondents' organizations are moderately satisfied with integration processes for accessing and integrating data from traditional legacy sources such as mainframes (18% very satisfied, 38% somewhat satisfied, and 20% needing a major upgrade) and enterprise applications (21%, 46%, and 21%, respectively). As organizations modernize data architectures, they should not overlook problems accessing data in legacy applications and mainframe systems. Research results suggest that many organizations are interested in automated solutions that would provide an upgrade to the legacy (often heavily manual) data integration they currently use to connect to older data sources.

Enterprise applications such as ERP have complex data structures that make data extraction and transformation slow and difficult. Data integration development costs and lack of skilled personnel can limit the types of reports and analytics. To provide complete data access sooner, some organizations replicate data from on-premises enterprise applications and mainframes to

cloud data storage. This enables organizations to consolidate application data more quickly into a cloud data lake or use a data pipeline to replicate it to a cloud data platform for transformation as needed for analytics workloads.

Trends such as SaaS adoption and digital transformation are accelerating the growth of data in the cloud. Many organizations today have a long-term strategy of migrating ERP and other business applications to SaaS systems. However, in practice, cloud migration often lands data on multiple cloud provider platforms. In Figure 10, just over half of respondents (53%) show satisfaction with access and integration of data from more than one cloud provider platform, with 16% very satisfied; 24% are looking for a major upgrade in technology to handle multicloud data access and integration.

Priorities for Supporting Analytics

To support growth in data visualization, analytics, and AI/ML, organizations need to address a number of issues. The most common shared priority is to streamline analysis of new data sets (42%) followed by improving analysis of data relationships across sources (41%; see Figure 11). Automation is vital to streamlining analysis at scale and improving speed as analytics grow more complex. Figure 11 shows that 38% say automating discovery of actionable data insights (e.g., for recommendations) is a top objective.

Forty-two percent of respondents say it is a priority to streamline analysis of new data sets; 41% want to improve analysis of data relationships across sources.

Frameworks such as DevOps and DataOps are valuable for helping organizations map out and orchestrate numerous pipelines and processes. Having a big-picture framework can be important as organizations add pipelines for real-time data streaming; 33% in Figure 11 say reducing data latency, such as with real-time data streaming, is a priority.

To modernize data integration for growth in visualization, analytics, and AI/ML, which of the following objectives are the highest priority for your organization? Select up to five.

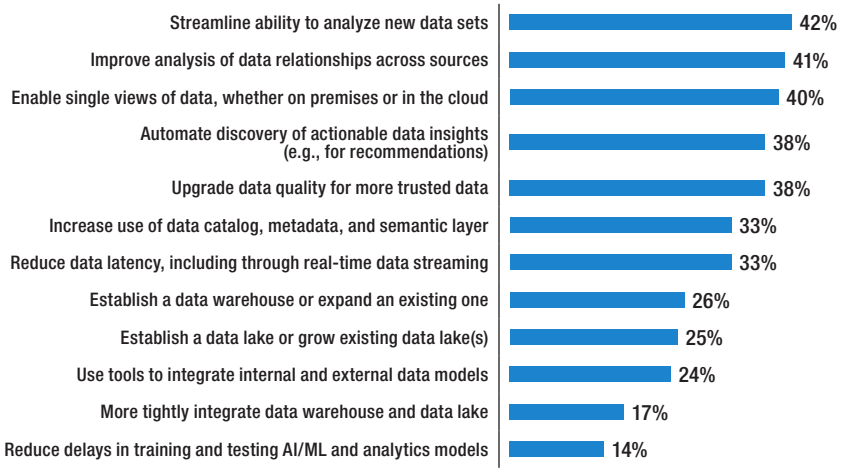


Figure 11. *Based on answers from 424 respondents.*

Many respondents (40%) prioritize enabling single views of data, whether on premises or in the cloud. To address this priority, organizations need to avoid letting new data platforms in the cloud become data silos that thwart single views. Depending on the use case, data virtualization could be a solution for avoiding heavy data movement involved in consolidation. However, other

scenarios demand that organizations consolidate data into a unified data architecture; 17% say it's important to more tightly integrate their data warehouse and data lake.

Improving Data Trust: Essential to Analytics Success

Data trust transcends technology. Users need to be confident in the data for visualization and analytics and what they share for consumption in reports, metrics, dashboards, and notifications. Data trust plays a central role in whether organizations realize business value from analytics when users share insights externally with partners, suppliers, and customers, including in a data marketplace or exchange.

Nearly half of research participants (45%) say their organizations are monitoring data quality to improve data trust.

From governance to data quality, organizations can take a variety of steps to improve data trust. Figure 12 displays which actions those surveyed regard as valuable. The largest percentage (45%) checked monitoring data quality. Depending on the use case—for example, financial performance management or exploratory analytics—organizations can apply a range of practices, data standards, and technologies to address quality problems in data creation, collection, transformation, analytics, visualization, and sharing.

To improve users' trust in the data, which of the following actions are being undertaken within your organization? Select all that apply.

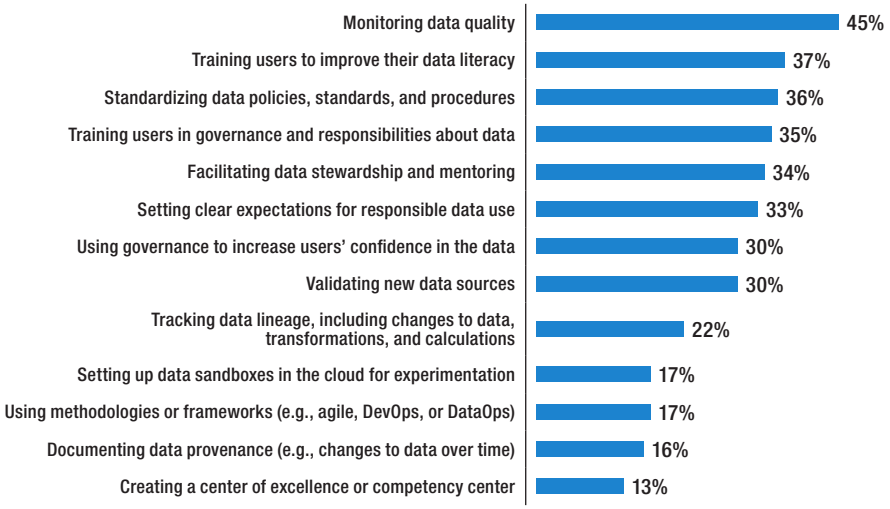


Figure 12. *Based on answers from 357 respondents.*

More than a third of respondents (37%) suggest that training users to improve data literacy would have value. Data literacy addresses human aspects of how people interact with data. The primary goal is to raise individuals' proficiency in understanding what data means and their ability to communicate and share analytics insights.

A second goal of data literacy is to increase people's responsibility and accountability for how they collect, integrate, prepare, and protect data. This is crucial to governance, adherence to data privacy regulations, and overall success in building trust in data and analytics through internal quality standards.

Training users in data governance and responsibilities is something that 35% of respondents regard as important to improving trust; 30% say using governance to increase users' confidence in the data is important. A significant percentage (36%) would standardize data policies,

standards, and procedures. Rounding out the top five (at 34%) is facilitating data stewardship and mentoring, which are important ingredients in both successful governance and a thriving analytics culture in which users share best practices and insights about data sets.

Cloud Data Integration and Management

For data integration and management, cloud computing offers scalability, flexibility, cost elasticity, and more rapid deployment. Organizations increasingly want modern data architectures centered in the cloud so data integration processes are closer to new data generation.

Thus, the focus today is on how to get data to the cloud faster and with less difficulty—and once in the cloud, how to manage it to realize business value and return on investment (ROI). Organizations use a range of cloud migration strategies. TDWI asked research participants which of the following types of migrations best describes their primary strategy (figure not shown; note that because many organizations use a mixture, respondents did not have to choose just one):

The focus today is less on justifying the move to the cloud and more on how to get there faster and with less difficulty—and how to realize greater business value and ROI.

- **Lift and shift.** This largest-scale approach configures a cloud-based image of existing on-premises data platforms. For example, organizations will keep the same data warehouse model using the same underlying software stack but deploy it on a cloud provider's platform. Just under one-fifth of respondents (18%) say this accurately describes their primary strategy; 39% say it is somewhat accurate and 22% say it is not accurate (21% don't know).
- **Phased migration**. With this approach, organizations migrate pieces incrementally. They test and validate components to enable users to move dependent data applications (e.g., BI reporting and analytics) to the cloud data platform. When planned well, a phased migration enables organizations to evaluate cloud services and align them with workloads. Just over one-fifth of respondents (22%) say this accurately describes their strategy; 46% say it is somewhat accurate and 13% say it is not accurate (19% don't know).
- **Start small and grow.** Organizations that have a cloud-first strategy for new systems often use this approach. For example, organizations might develop new (but limited) reporting and analytics using cloud services to see if those services are the right choice before they scale up. They can also evaluate cost drivers as they add users and increase the volume of data moving across networks and stored in the cloud. The highest percentage surveyed say this accurately describes their strategy (29%) and even more say this is somewhat accurate (40%); just 16% say it is not accurate (15% don't know).
- **Leave on-premises systems in place but anything new goes to the cloud.** This variation accurately describes the strategy for 28% of respondents' organizations with 27% calling it somewhat accurate and 24% saying it is not accurate (20% don't know). Organizations using this strategy must plan on managing data integration, access, and governance across a hybrid of on-premises and cloud systems, often with multiple cloud providers (multicloud).
- **No overall strategy.** In this case, cloud migration is piecemeal and driven by project teams' immediate and specific needs. Just 13% say this is an accurate description of their strategy, although 32% say it is somewhat accurate; 29% say it is not accurate and 25% don't know.

Top Objectives for Cloud Data Integration and Management

Cost flexibility is top of mind for cloud data integration and management; organizations need flexibility to meet unanticipated needs.

TDWI finds that most organizations have multiple drivers behind cloud data integration and management. In Figure 13, we can see which objectives are most prevalent among respondents' organizations. Cost flexibility is top of mind; 43% selected expanding data access and integration with lower up-front investment. With the cloud, organizations can gain business agility; they can align data integration and management investment with business operational needs rather than having to follow budgeted IT capital expense restrictions.

Which of the following objectives are most important in guiding your organization's strategy for cloud data integration (DI) and management? Select up to five.

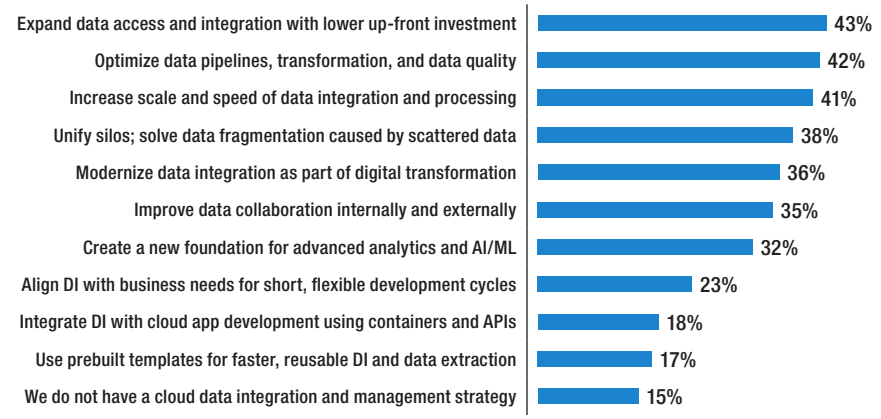


Figure 13. *Based on answers from 396 respondents.*

Nearly one-quarter (23%) say that using the cloud to align data integration with business needs for short, flexible development cycles is valuable. Some respondents point to the use of prebuilt templates for faster, reusable data integration and extraction (17%); this can reduce development time for repeatable workloads. Over one-third (35%) note that they view cloud data integration and management as a way to improve data collaboration internally and externally.

Forty-two percent of respondents indicate that optimizing data pipelines, transformation, and data quality is a key objective. This suggests that organizations want to use scalable platforms in the cloud to drive faster performance on larger data volumes.

About the same percentage (41%) say that increasing the scale and speed of data integration and processing is an objective guiding their cloud data integration and management. A significant percentage (38%) of respondents' organizations regard cloud migration as an opportunity to unify silos and solve data fragmentation, a step that can reduce latency.

Advanced analytics expansion drives some cloud data integration strategies. Figure 13 shows that 32% of respondents say their organizations want to create a new foundation for advanced analytics and AI/ML. Today, organizations can easily set up cloud data lakes using object storage on cloud platforms. This offers a faster path to launching data science projects involving exploratory, predictive, and real-time analytics. Data scientists can develop and test AI/ML models and algorithms that depend on access to hundreds of terabytes (if not petabytes) of data contained in the data lake and sourced from log files, social media, mobile device data, and more.

Top Challenges with Cloud Data Integration and Management

Respondents highlight a number of challenges in migrating their data integration and management to the cloud. In Figure 14, we can see some primary concerns. Three key areas include:

- **Cost concerns.** More than a third (36%) cite data ingestion, egress, extraction, and migration costs as a challenge. Cloud data egress costs, which can rise rapidly when an organization moves volumes of data out of one cloud provider's platform to another location, are a widespread concern. Nearly a third of respondents (32%) indicate concerns about lack of visibility and metrics for managing costs and 31% say they have higher than expected costs for user data access and interaction. One-quarter (25%) cite constraints on scalability and the number of users they can support concurrently due to cost concerns.
- **Problems with speed to insight and access to real-time data.** Nearly a third (32%) say cloud data integration and management are not providing data insights as fast as they are needed. A number of factors affect speed to insight, including not getting data loaded into target data platforms for users fast enough. However, for 30% of organizations, gaining better access to live or real-time data is a challenge. One-quarter of organizations (25%) are finding that data silos are increasing due to poorly planned cloud adoption. The inability to gain a complete view of or access to disparate data is often a factor that slows speed to insight.
- **Governance and data trust challenges.** A significant percentage of organizations (29%) note concerns about access control, security, and governance challenges. As organizations augment on-premises data platforms with new cloud services or are in the middle of data platform migrations, they need to expand governance oversight. About a quarter of respondents (24%) say lagging data trust due to quality and consistency issues is also a challenge. With new data types and higher volumes of data housed in cloud data storage, organizations need to ensure that data quality rules and standards fit the new environment.

Which of the following issues are the most challenging regarding your organization's migration to and use of the cloud for data integration and management? Select up to five.

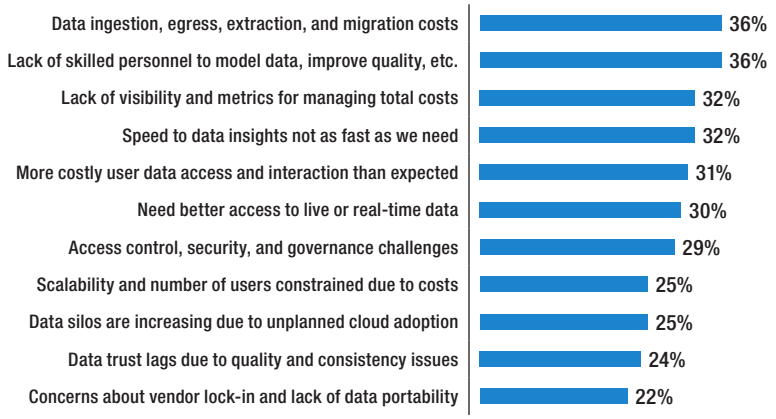


Figure 14. *Based on answers from 387 respondents.*

Data Architecture and Future Data Strategies

Data architecture brings together strategies, practices, technologies, and, increasingly, cloud services. Technologies include data storage, computational processing, and related tools for data collection, movement, and integration. With today's varied users and workloads, organizations cannot stand pat with data strategies and architectures constructed for an earlier age—defined by limited BI reporting, few data science projects, and monolithic data warehouses collecting data from a small range of data sources. New technologies, services, and practices enable organizations to support a spectrum of data-driven use cases.

To close out our examination of research data, we look at organizations' satisfaction in achieving objectives with their overall data architectures and how participants rate the importance of key data and information integration trends shaping data strategies.

Organizations show relative satisfaction with a variety of options, such as change data capture, for improving data refresh rates and reducing latency.

Research shows moderate satisfaction with data refresh rates and search and query speed. We asked research participants about their organizations' satisfaction with how their overall data architectures (including the data warehouse, data lake, and related data integration processes) achieved common objectives. Results indicate the most satisfaction with how they provide users with appropriately refreshed data: 57% are satisfied (including 15% very satisfied); 27% are unsatisfied and 16% don't know (figures not shown). Users need data refreshes so historical values in the data warehouse or analytics platform are in sync with source data values. Change data capture replication can be helpful for keeping track of changes and updating users' applications, especially when data is changing continuously.

The second-highest area of satisfaction reported is improving the speed of data searches and queries. Just over half (53%) are satisfied with their data architecture for this objective and 30% are unsatisfied (17% don't know). This suggests that satisfied organizations have skilled personnel who know how to write faster and more efficient SQL queries and searches as well as platforms or cloud services with adequate processing power and query optimization. Many respondents' organizations need data architecture improvements to handle advanced analytics, AI/ML, and predictive modeling; 34% are satisfied with how their architecture handles this objective, but only 9% are very satisfied, and 38% are unsatisfied.

Just under half are satisfied with data centralization and integration of their data warehouse and data lake. Many organizations have strategies for consolidating fragmented data using centralized databases or storage in the cloud. In our research, 48% of organizations say their data architecture is satisfactorily centralizing data to reduce silos and fragmentation, although only 11% are very satisfied; 34% are unsatisfied and 18% don't know. Similar percentages report satisfaction with integration between the data warehouse and data lake; 47% are satisfied with 10% very satisfied, 26% are unsatisfied, and 27% don't know, which may include respondents who do not have both a data lake and a data warehouse. As discussed earlier, a unified data lakehouse tightens integration between a data lake and data warehouse and supports data strategies for consolidating disparate data silos.

Organizations are using data virtualization to provide faster and easier access to data outside the data warehouse.

About two in five are satisfied with data virtualization to complement the data warehouse. A data virtualization layer combines data from multiple sources to create new logical views. Users can then query the views from within self-service BI and analytics tools or applications without having to know how to access the data at the sources. The research indicates that a significant percentage of respondents (43%) are satisfied with how they are using data virtualization layers in tandem with their data warehouses. However, 31% are not satisfied with this strategy and 26% don't know.

TDWI additionally found that 48% of respondents' organizations are satisfied with their data architecture for shielding users from the complexity of accessing data and getting results; 34% are not satisfied and 18% don't know. Semantic data integration layers and data virtualization provide technologies for achieving this objective.

Organizations exhibit moderate satisfaction with multimodel support. The growing spectrum of workloads, particularly for data science, puts pressure on data architectures to be multimodel—that is, to provide unified support for both relational and nonrelational data models such as NoSQL databases. The goal of a unified, multimodel data architecture is to integrate all types of processing, including classic BI reporting and OLAP, real-time streaming analytics, and advanced analytics such as AI/ML. About two in five respondents (42%) say their organizations are satisfied with current multimodel support for different data models; 30% are dissatisfied and 29% don't know.

Current and Future Data Strategy Objectives

Data and information integration are central to how organizations deepen understanding of subjects important to the business, including customer behavior and market trends, supply chain disruptions, risk exposure, new product and service opportunities, and competitive behavior. Organizations need to continuously modernize data strategies to serve evolving business requirements and take advantage of technology innovation.

Organizations need to continuously modernize data strategies to serve evolving business requirements and take advantage of technology innovation.

TDWI asked research participants about the importance of a series of objectives to their current and future data strategies (see Figure 15). The focus is on objectives for better data strategy support for broader and deeper analytics, effective use of knowledge about the data, and incorporation of smarter and more automated modern technologies.

How important to your organization's current and future data strategy are each of the following objectives?

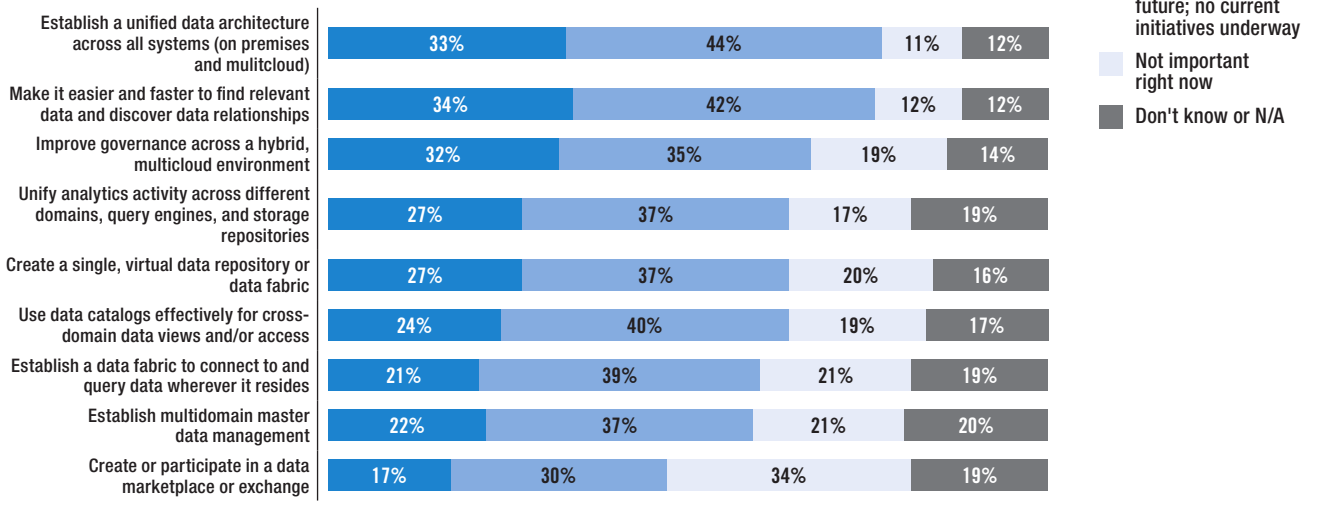


Figure 15. *Based on answers from 358 respondents. Ordered by highest combined "very satisfied; part of our current strategy" and "important for the future; no current initiatives underway" responses.*

Most want to establish a unified data architecture across all systems (on premises and multicloud). More than three-quarters of respondents (77%) say it is a priority to have one architecture supporting all types of workloads rather than disparate and disconnected systems. One-third (33%) say it is very important and part of their current strategy and 44% say it is important for the future but they have no current initiatives underway to unify all systems in their data architecture.

Organizations want to make it easier and faster to find relevant data and discover data relationships. Figure 15 shows that 76% view this objective as important, with 34% indicating it is very important and part of their current strategy. We saw earlier that respondents have relatively strong satisfaction with visualization and analysis of data relationships in their organizations. However, as users, applications, and AI/ML programs need more diverse data types and sources, they require new technologies and practices. These include knowledge graphs, graph databases, and semantic data integration.

Research participants indicate that a data fabric is important to their current and future data and analytics strategies.

Interest is strong in data fabrics and meshes. Research participants indicate that a data fabric is important to their current and future data and analytics strategies. With many organizations living in hybrid and elastic multicloud data environments, data fabric and data mesh solutions are attracting attention as ways to overcome complexity.

A data fabric uses the concept of a broad but connected network of information to unify environments and make it seamless to view, access, and manage all data. It applies semantic richness based on metadata and additional knowledge assets; fabrics feature automated intelligence to speed discovery of data relationships and support advanced analytics. A data virtualization layer or virtual repository can be an important foundation for a data fabric. A data mesh architecture has similar attributes, but leaves more control with decentralized domains that hold copies of data rather than the fabric approach of creating a single, central physical or virtual data repository.

Respondents show strong interest in data fabric and data mesh approaches for addressing challenges in hybrid multicloud environments that are often changing. Nearly two-thirds of respondents' organizations (64%) say it is important to create a single virtual data repository or data fabric, with 27% indicating that it is very important and part of their current strategy.

The same percentages hold for the importance of unifying analytics activity across different domains, query engines, and storage repositories (64% important, with 27% indicating it is very important and part of their current strategy). In addition, a strong percentage of respondents find it important to establish a data fabric to connect to and query data wherever it resides; 21% say this is very important and part of their current strategy and 39% say it is important for the future but have no current initiatives.

Two-thirds want to improve governance across hybrid multicloud environments. Data governance is challenging enough in traditional environments, but organizations today need to ensure adherence to governance rules, regulatory policies, and quality processes across hybrid multicloud environments. A significant percentage of respondents (67%) say they need to improve governance so it covers hybrid multicloud environments. Nearly one-third (32%) call this part of their current strategy, and 35% see it as important for the future. Data fabric governance services, combined with metadata resources, could help organizations securely control authorized access to data sources and ensure adherence to governance policies.

Organizations want data catalogs and MDM for handling multiple domains. Most organizations endorse the importance of using data catalogs effectively for cross-domain data

views and/or access; 24% say it is very important and part of their current strategy, and 40% say it is important to the future although they have no current initiatives underway. Data catalogs that offer cross-domain metadata intelligence about data's relevance, quality, lineage, and governance constraints could improve collaboration on data across domains such as different departments or business processes and open up new analytics perspectives.

Just under three in five respondents (59%) say it is important to establish multidomain MDM, with 22% saying it is very important and part of their current strategy. Expanding from single- to multidomain MDM enables organizations to more efficiently manage master data across diverse domains. The multidomain MDM trend fits with movement toward data fabrics and data mesh architectures. Multidomain MDM can provide standardized master data definitions coordinated with business entities to reduce confusion among internal and external users, applications, and automated processes.

Nearly half say a data marketplace or exchange is important. Organizations are showing interest in data services offered through storefronts in a data marketplace or exchange. These are curated spaces for buying, selling, and sharing data and analytics. To gain fuller operational visibility and to support data-hungry analytics, organizations are looking for a faster path to incorporate new data.

A data marketplace or exchange offers a curated space for buying, selling, and sharing data and analytics; 17% say it is very important to their data strategy.

A data marketplace or exchange can provide a forum for data monetization opportunities. Some organizations are using the approach for internal data services to make it easier for users to share and locate relevant data sets. Nearly one-fifth of respondents (17%) say that creating a data marketplace or exchange is very important and part of their current strategy and 30% say it is important for the future; 34% say it is not important right now.

Recommendations

To conclude this report, here are 10 recommendations for developing strategies to deliver greater business value from data integration, data catalogs, and related technologies and platforms.

Address latency in data pipelines and transformation processes. Fresher data and quicker updates contribute to timelier decisions and more efficient business operations. 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. However, there is no single answer; different user requirements demand different solutions. Organizations need multiple options for reducing latency. Our research shows interest in reducing data movement through data virtualization and ELT data pipelines. Many organizations want to take advantage of technologies for real-time data streaming. Organizations also need seamless end-to-end integration of processes within data pipelines.

Focus on data integration to accelerate benefits of cloud migration. Whether to augment on-premises systems or as part of a cloud-first strategy, organizations are moving significant data management to the cloud. With reducing costs and enabling scalability with lower up-front investment as major drivers, organizations should not ignore data integration's importance. Our research finds that organizations want better visibility into data integration processes so they can increase efficiency and reduce costs. Organizations seeking faster speed to insight need to modernize data integration, reduce data fragmentation due to too many data silos in the cloud, and update governance policies.

Address challenges in hybrid multicloud environments. Many organizations have some systems situated on premises and others on multiple cloud provider platforms. Organizations may choose a multicloud strategy to take advantage of the different strengths of each platform and avoid vendor lock-in. However, this adds complexity. This report's research shows that organizations are interested in establishing a unified data architecture across all systems. Data virtualization, data fabrics, and semantic layers built on cross-domain data catalogs and MDM are among the solutions organizations should evaluate.

Modernize data knowledge to enhance data pipelines, preparation, and integration. Having easily accessible and up-to-date knowledge about data, its lineage, its location, and its relationships is valuable. It shortens the time it takes people to find, prepare, and use data. Organizations can reduce confusion about data quality and improve collaboration and governance. These virtues are important as organizations develop numerous pipelines and integrate different data types and sources. Modern data catalogs are automating previously manual processes and employing AI/ML for better metadata accuracy and scalability. Semantic tools such as knowledge graphs capture deeper information about how data sets are used and related. These solutions contribute to faster and more satisfactory data pipelines, preparation, and integration.

Consolidate data silos and tighten data warehouse and data lake integration. With different types of users and workloads requiring access to diverse data, organizations need to make it easier to discover, view, access, govern, and manage all their data. Modern data architectures should encompass relational and NoSQL systems such as document and graph databases to tap the right system for different requirements. Organizations should evaluate the benefits of tightening data warehouse and data lake integration in the cloud.

Evaluate data virtualization for delivering transparent data access and reducing data movement. A data virtualization layer can shield users from having to know how to access the data or where the data is stored. Organizations should evaluate data virtualization for reducing data movement and optimizing query processing at the sources to take advantage of local processing power. Organizations should take steps to tighten integration between data catalogs, semantic data layers, and knowledge graphs with data virtualization layers.

Evaluate data fabrics and data mesh architecture. Limited data access is a problem when organizations need timely, complete views of all relevant data about customers, supply chains, business performance, public health, and more. Research in this report shows strong interest in the notion of a data fabric or data mesh architecture. The notion is becoming important as organizations put more data in multiple cloud-based storage platforms. Organizations should evaluate data fabrics and data mesh architectures for improving response in dynamic situations where your organization needs to perform analytics right away on data in multiple sources. Data fabrics and data mesh architectures typically depend on metadata, knowledge graphs, and semantic data knowledge, potentially integrated with a data virtualization layer.

Modernize data integration to improve analytics and discovery of data relationships. This report's research shows that analytics projects are an important driver behind data integration modernization. Today, analytics and AI/ML projects demand both breadth and depth of access to data. Visualizing and analyzing data relationships is important, according to this research. Knowledge graphs and graph databases can potentially help organizations discover and analyze complex data relationships faster and more accurately.

Ease self-service data preparation, integration, and transformation. Easier-to-use functionality enables non-IT users to do more on their own. This is vital as organizations seek to empower users to discover data insights and use them effectively to make better decisions, engage in informed customer relationships, and bring efficiency to business processes. However, organizations do not want users mired in data preparation, pipeline development, and data transformation. It is important to push toward self-service data access, expanded use of metadata catalogs, and automated data pipeline development. Organizations should also consider knowledge graphs and graph databases to reduce complexity in data preparation, pipeline development, and data transformation for advanced analytics.

Address gaps in data governance and improve data trust. Organizations that have set up governance to fit traditional data warehousing need to update rules and policies as environments expand into diverse data stores on hybrid multicloud platforms. This research shows that improved governance is a top priority as organizations modernize data catalogs and data integration strategies. Knowledge graphs, data fabrics, and data mesh architectures can play important roles in adapting governance to cover whole data environments. Governance and associated data stewardship are critical to improving data trust. Data quality is a priority for data trust; organizations need to evaluate solutions that scale data quality and improve confidence that the data is right.

actian.com

Actian, the hybrid data management, analytics and integration company, delivers data as a competitive advantage to thousands of customers worldwide. Through the deployment of innovative hybrid data technologies and solutions Actian ensures that business critical systems can transact and integrate at their very best – on premise, in the cloud or both. Thousands of forward-thinking organizations around the globe trust Actian to help them solve the toughest data challenges to transform how they run their businesses, today and in the future. For more, visit https://www.actian.com. "Actian", "Avalanche" and "Activate your Data" are trademarks of Actian Corporation and its subsidiaries. All other trademarks, trade names, service marks, and logos referenced herein belong to their respective companies.

denodo.com

Denodo is a leader in data virtualization providing agile, high-performance data integration, data abstraction, and real-time data services across a broad range of enterprise, cloud, big data, and unstructured data sources at half the cost of traditional approaches. Denodo's customers across every major industry have gained significant business agility and ROI by enabling faster and easier access to unified business information for agile BI, big data analytics, web and cloud integration, single-view applications, and enterprise data services.

The Denodo Platform offers the broadest access to structured and unstructured data residing in enterprise, big data, and cloud sources, in both batch and real time, exceeding the performance needs of data-intensive organizations for both analytical and operational use cases, delivered in a much shorter time frame than traditional data integration tools.

The Denodo Platform drives agility, faster time to market, and increased customer engagement by delivering a single view of the customer and operational efficiency from real-time business intelligence and self-serviceability.

For more information visit www.denodo.com, follow Denodo via twitter@denodo, or contact us to request an evaluation copy at info@denodo.com.

snaplogic.com

SnapLogic is a leader in self-service application and data integration. Our solutions make it fast and easy for you to connect applications and data across the enterprise so you can improve business processes, accelerate decision making, and drive better business outcomes. The SnapLogic Intelligent Integration Platform accelerates data and process flow across applications, databases, data warehouses, big data streams, and IoT deployments—whether on premises or in the cloud. Unlike traditional integration software that requires painstaking, hand-crafted coding by teams of developers, SnapLogic's simple but powerful platform enables both IT and business users to create quality, scalable data pipelines that get the right data to the right people at the right time.

www.stardog.com

Stardog, a leading enterprise knowledge graph platform, turns data into knowledge to power more effective digital transformations. Industry leaders including BNY Mellon, Bosch, and NASA use Stardog to create a flexible data layer that can support countless applications. With Stardog, customers reduce data preparation timelines by up to 90%. Stardog is a privately held, venture-backed company headquartered in Arlington, VA. For more information, please visit www.stardog.com.

trifacta.com

Trifacta delivers an intelligent, collaborative, self-service data engineering cloud platform to transform data, ensure quality, and automate data pipelines, enabling consumable data at any scale. Data engineers, analysts, and scientists can collaborate to produce useful data for advanced insights and analytics. With universal data connectivity, Trifacta provides flexibility to connect to any data from any source for any application. Using an AI-assisted, self-service approach, Trifacta users can collaborate to evaluate, correct, and validate data quality, accelerate data transformation, and automate robust data pipelines. Trifacta's open cloud platform drives increased business value through informed decisions.

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.

Transforming Data With Intelligence™

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

E info@tdwi.org

tdwi.org