research

BEST PRACTICES REPORT

Q1 2020

# Faster Insights from Faster Data

Technologies and practices for better speed to insight

By David Stodder

Co-sponsored by:

denodo

tdwi
Transforming Data
With Intelligence™

research

BEST PRACTICES REPORT

Q1 2020

# Faster Insights from Faster Data

Technologies and practices for better speed to insight

By David Stodder

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

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

## 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 Research

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.

## 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), Philip Russom (prussom@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 diligently reviewed outlines, survey questions, and report drafts. Finally, we would like to recognize TDWI's production team: James Powell, Peter Considine, Lindsay Stares, and Rod Gosser.

## Sponsors

Ascend.io, Denodo, Matillion, SAS, and Wyn Enterprise by GrapeCity sponsored the research and writing of this report.

# Research Methodology and Demographics

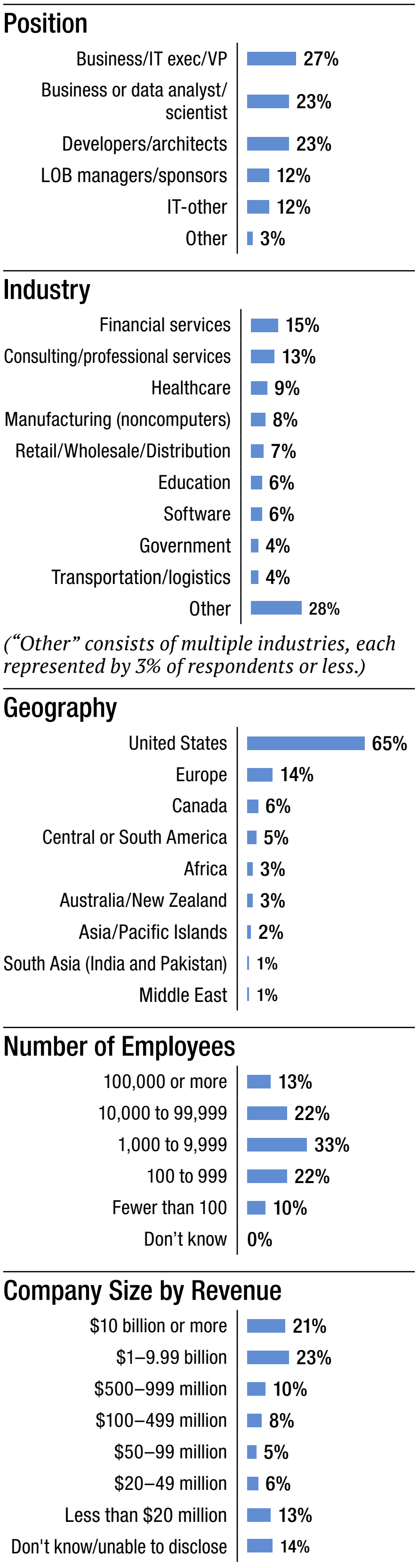
**Report Purpose.** Speed to insight has long been an important objective, but users are frustrated by delays in getting from data to insights, sometimes due not to technology but to project disarray. Challenges are only growing as data gets more voluminous, varied, and faster with streaming data. Fortunately, solutions are providing options to fit ambitious use cases and analytics and AI workloads. This report examines the challenges and offers recommendations.

**Survey Methodology.** In September 2019, TDWI sent an invitation 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 147 respondents. Not all respondents completed every single question; however, all responses are valuable and so 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.** Just over one-quarter of survey respondents are business or IT executives and VPs (27%). The second-largest percentages are business or data analysts and data scientists (23%) and developers and data, application, or enterprise architects (23%). Line-of-business (LOB) managers and business sponsors account for 12% of the respondent population. "IT-other" titles account for the same percentage (12%).

Financial services is the largest industry group (15%), followed by consulting and professional services (13%), healthcare (9%), manufacturing (noncomputers) (8%), and retail, wholesale, and distribution (7%). Respondents from education and software industries each accounted for 6%, and government respondents for 4%, the same as from transportation and logistics (4%). The remaining 28% are from a variety of other industries. Nearly two-thirds of survey respondents reside in the U.S. (65%), with Europe (14%) and other regions following. Respondents come from enterprises of all sizes.



Position

Industry

*("Other" consists of multiple industries, each represented by 3% of respondents or less.)*

Geography

Number of Employees

Company Size by Revenue

*Demographics based on 147 respondents.*

# Executive Summary

This TDWI Best Practices Report shows strong interest in achieving faster insights to seize business opportunities, but points out where organizations need to address technology and practice weaknesses.

Organizations today place a high priority on fact-based, data-driven decision making. This makes speed to insight a competitive advantage. If a retailer can analyze data to uncover a trend in customer preferences before other retailers in the marketplace, they can gain an edge, potentially delivering higher market share, customer loyalty, and profitability. If an insurer or government healthcare agency can use predictive models to detect a fraud scheme before it has a chance to do significant damage, it can save costs and avoid public embarrassment.

There are many more cases where faster insights can be beneficial. However, achieving faster insights can only happen if delays and bottlenecks that exist throughout data life cycles are addressed using better practices and modern technologies. This TDWI Best Practices Report examines where organizations are coming up against barriers to getting relevant data from sources into the right condition for analytics, for developing artificial intelligence (AI) programs such as machine learning to discover insights, and for delivery to the array of users who need insights in time to solve business problems.

Some of the challenges relate to how organizations put together project development teams and determine deliverables. Setting project objectives is a challenge; TDWI research finds that just 9% of organizations surveyed regard themselves as very successful in identifying value measures and quantifiable objectives (see Figure 4 in this report). As projects move toward deliverables, less than half of those surveyed regard their organizations as either good or excellent at testing prototypes and developing proofs of concept. To address these weaknesses, organizations are implementing agile, DataOps, and other methods to help them better organize projects and move faster to create value.

Technology advances are also key. This report discusses how organizations can reduce latency in data preparation, transformation, and development of data pipelines. It details how organizations could be using data catalogs, metadata repositories, and data virtualization more effectively, including for governance. With expense and scalability identified by research participants as two main issues they face, it is not surprising that cloud-based data management, integration, transformation, and development are popular. This report discusses how moving to the cloud solves some problems but spotlights other issues, such as governance and finding the right balance between centralization and self-service environments.

Data itself is getting faster as organizations begin to analyze new sources including streaming data coming from sensors, websites, mobile devices, geolocations, and more. The report finds that some organizations use streaming, real-time analytics and AI to automate decisions and deliver actionable recommendations to users. TDWI recommends that organizations focus on well-defined objectives and also devote attention to their big-picture strategy to avoid letting complexity slow innovation.

# Speed to Insight: About More than Just "Fast"

How important is it to your organization to invest in solutions, cloud services, and practices that can enable faster analytics and data consumption and faster data integration, preparation, transformation, and processing? The answer we received to this introductory survey question was clear: 80% of respondents say it is important, with 39% indicating it is "extremely" important (figure not shown). Getting faster is an objective shared by just about all organizations.

Nearly everyone agrees that if business executives, managers, and frontline personnel do not have to wait for the most current insights—or be forced to use yesterday's or last month's information when they really need the very latest—they could make more timely decisions, seize fleeting opportunities, and serve customers and partners more effectively. Yet, the rub is that faster is not better if the data and information are not accurate, complete, or fit for the purpose.

To be sure, for some use cases such as data science, exploratory analytics, and AI, the faster that users and AI programs can get access to raw, live data that's just been recorded or streamed in real time, the better. However, for most users and applications, our research finds that data quality, accuracy, completeness, and relevance are more important than just pure speed. Users want to know if they can trust the data; they often need to know where it came from, how it relates to other data, and how it has been transformed and aggregated. Before the data can flow, organizations also need to ensure that it is secure and governed in accord with regulations.

This TDWI Best Practices Report examines experiences, practices, and technology trends that focus on identifying bottlenecks and latencies in the data's life cycle, from sourcing and collection to delivery to users, applications, and AI programs for analysis, visualization, and sharing. Reducing time to insight depends on applying technologies and appropriate practices that improve matters at each phase in the life cycle. Organizations must take into account the type of user and their context; "fast" and "real time" can have different meanings depending on these factors. For some, just getting the data or insights at the time that they need it is the equivalent of real time (what some used to call "right time"); for others, real time means reducing latency to the smallest possible interval between the data's creation and its availability for analysis and visualization.

As technologies advance, including through opportunities created by cloud computing, traditional ways of working with data need to be reconsidered. This report looks at organizations' level of interest in innovative technologies such as AI, data streaming, and real-time analytics as well as newer ways of integrating, processing, preparing, and transforming data. AI is a change agent in all these phases as well as the user experience itself.

Through AI-driven augmentation, BI and analytics solutions are evolving to offer  recommendations about data sets users might explore, visualizations to use, and, ultimately, decisions or actions to take. Such recommendations can result in faster insights, which can translate into more responsive and proactive engagement with customers and partners as well as strategies for success in the marketplace, supply chains, and other business contexts. With the appropriate stack of technologies to support it, users can experience right-time or actual real-time dashboards in contexts that focus attention on particular key performance indicators and other metrics. These can be supplemented with prescriptive, AI-driven recommendations based on the data.

AI and advanced analytics on top of streamed, real-time data feeds enable organizations to spot trends and patterns, apply predictive insights, and potentially automate responses to situations, particularly those in fraud detection, securities trading, e-commerce, emergency healthcare, and population health where instantaneous response is critical. Solutions that employ AI techniques such as machine learning can help organizations profile, transform, and enrich real-time data as it flows through pipelines so these steps are faster and more efficient than with traditional extract, transform, and load (ETL) technologies.

AI techniques are being combined with other technologies to accelerate data processing, access, and interaction. AI is becoming useful for governance by enabling organizations to learn more, faster about new big data and improving tracking to ensure that data use adheres to governance rules and policies. AI's contribution to data lineage tracking can help with other data management requirements such as performance and concurrency that demand monitoring of what data is being used and shared in preparation and pipeline processes.

## Data Literacy and the Analytics Culture

Organizations need to build cultures that nurture good practices and support data-informed decision making. They also need to raise data literacy so personnel are confident in their abilities to work with and share data and analytics.

Faster data flows and speedier data processing and transformation are critical, but organizations also need to develop a supportive and innovative culture for the data riches to have a meaningful effect. This involves establishing an analytics culture, which is about fostering leadership, communication, and collaboration to overcome resistance and build trust in the process of creating insights from data and applying them to decisions and actions. If the culture does not thrive, successes in analytics will remain isolated and not lift the entire organization's data-driven intelligence.

An analytics culture nurtures good practices in using data and analytics to support decisions throughout the organization. It supports challenging assumptions, learning from the data, and owning the outcome by objectively measuring the impact of decisions based on the analytics. It is not something that can be created overnight. Organizations need to work on developing an analytics culture that matures in a healthy direction over time.

Analytics cultures depend on raising the data literacy of individuals in the organization. Users may have great BI and analytics tools, but if they struggle to understand data, visualizations, and analytics and cannot effectively interact with and share data and insights, the tools will not be enough. As part of our research for this report, TDWI asked participants how they would rate the overall data literacy of users in their organizations (see Figure 1). The highest percentage say their organizations are "about average" (42%) in terms of users' ability to consume, analyze, interact with, share, and discuss data in the course of carrying out their roles and responsibilities. Nearly the same percentages rate their organizations' data literacy as "somewhat high" (22%) and "somewhat low" (24%).

**How would you rate the overall data literacy of users in your organization: that is, their ability to consume, analyze, interact with, share, and discuss data in the course of carrying out their roles and responsibilities?**

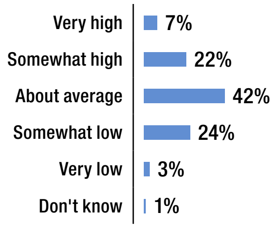


***Figure 1.*** *Based on answers from 147 respondents.*

Growth in data literacy can increase an organization's overall speed to insight and accelerate innovation with data. Personnel will be better prepared to contribute to projects for development of analytics, data-driven applications, and AI. This leads us into the first section of research results, which focuses on how organizations put together leadership, teams, and project development processes. Along with data literacy, these issues can either propel or thwart an organization's progress toward building value from data.

# Leadership, Organizational, and Project Challenges

"It's not just the technology, it's the people." At TDWI we hear this often. Adoption and implementation of the latest technologies will always be critical to faster and more effective data-driven decision making. However, TDWI research has long found that organizations fall short of realizing the value of technology investments due to organizational and project management difficulties. Key issues in these areas include strategy, culture, leadership, skills, and funding. Because most analytics and data management projects impact a range of stakeholders—from executive leadership and line-of-business (LOB) managers to IT developers, data scientists, business and data analysts, and frontline users—developing good collaboration and communication among them is vital.

Organizations fall short of realizing the value of technology investments for faster insights due to difficulties in strategy, culture, leadership, skills, and funding.

For this TDWI Best Practices Report, we asked research participants to rate how well stakeholders in their organizations collaborate to accomplish key steps in the life cycles of BI, analytics, AI, and data integration and management projects (Figure 2). Only a small percentage gave their organizations "excellent" ratings for any of the steps. Organizations surveyed appear strongest in identifying relevant data sources (65% combined excellent and good ratings) and identifying opportunities to achieve business benefits (61% combined). This suggests a relationship between the two: that is, stakeholders believe that if they can access and analyze certain data sources properly, it will lead to better business outcomes.

How would you rate your organization's collaboration between business stakeholders, IT, developers, data scientists, and analysts to accomplish the following steps in the life cycle of projects for BI, analytics, AI, and data integration and management?

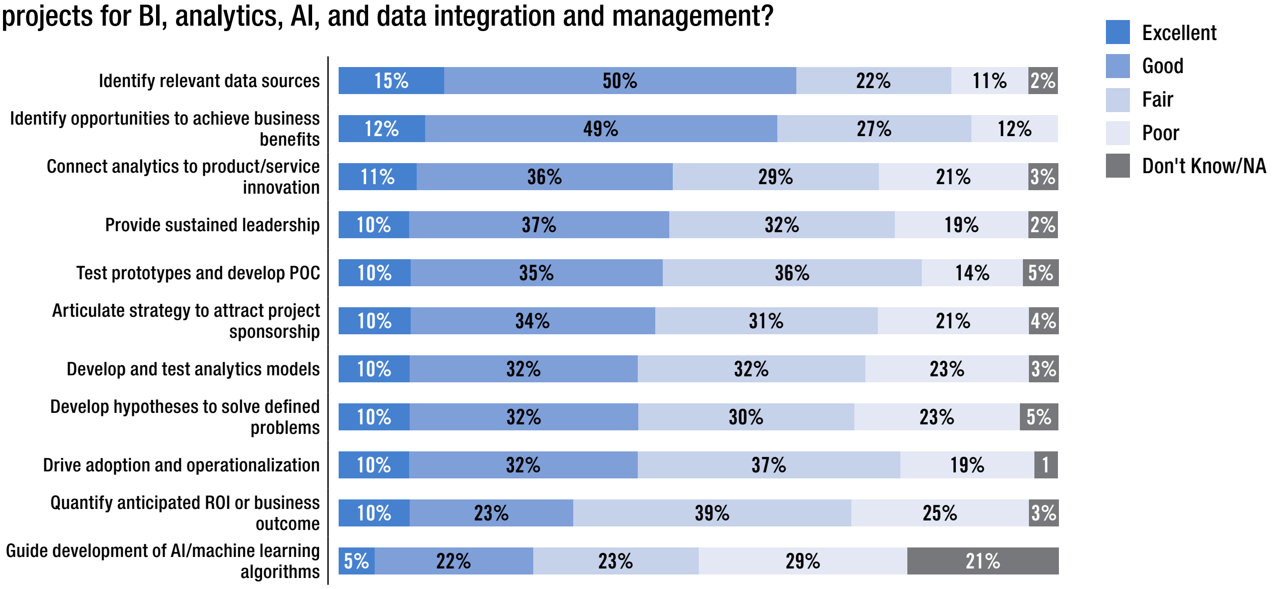


*Figure 2. Based on answers from 147 respondents. Ordered by highest combined "excellent" and "good" responses.*

Significant but smaller percentages of respondents gave their organizations good or excellent ratings for providing sustained leadership (47% combined good and excellent) and articulating strategy to attract project sponsorship (44% combined). Concerted leadership is key to keeping projects going and dealing with both short-term needs and long-term objectives. Project teams need to include personnel who can interpret business needs and articulate strategic vision so leaders sustain sponsorship of projects and provide appropriate funding. Leadership often asks teams to quantify anticipated return on investment (ROI) or business outcome; however, only 33% of research participants view their organizations as either good or excellent in doing this.

**Need for collaboration on analytics development.** The majority of organizations appear to have difficulty fostering collaboration on important steps in developing analytics. Less than half regard their testing of prototypes and developing proofs of concept (POCs) as either good or excellent (45% combined "excellent" and "good" answers) and slightly fewer give high marks to collaboration on development and testing of analytics models (42% combined). More than half of respondents give their organizations low marks on developing hypotheses to solve defined problems, which is important to connecting projects to specific issues where the data insights from analytics and AI are needed and can be tested; 30% regard collaboration on this step as "fair" and 23% indicate that it is "poor."

The majority of organizations struggle with getting projects to the point of operationalization. Less than half see collaboration to drive adoption and operationalization as a strength, with only 10% checking "excellent" and 32% calling it "good."

**Length of development before getting to actionable results varies.** In 2017, TDWI asked research participants about the length of their development, testing, and deployment cycles. We asked the same question for this report to see if cycles were any faster or slower. The results are fairly consistent between then and now; the fastest 15% say their cycles are two weeks or less, which is the same percentage as in 2017; in both, the slowest 20% say their cycles take more than four months. Both then and now, 17% of research participants say their cycles take 15–30 days. For cycles that take beyond 30 days, the comparison shows that organizations are becoming a little faster, with more in the 31–60 day range than previously (18% for this survey compared to 15% in 2017). Overall, however, it appears that the time it takes for development, deployment, and testing cycles to finish has not appreciably changed.

One factor that may help reduce cycle times is the use of agile, DevOps, and DataOps methodologies, which will be discussed in greater length in a moment. When we filter the results to look only at organizations that are using these methodologies, we see that development cycle times are less varied across the spectrum. The bulk of respondents' cycles are under 90 days, with 34% of this selected group completing cycles in 30 days or less.

## Methods and Frameworks for Increasing Speed and Quality

**Agile, DevOps, DataOps, and other alternative methods to "waterfall" development are popular and can help organizations deliver value sooner.**

A strong trend in recent years is the adoption of software development and engineering methods for BI, analytics, and data management projects. Although still appropriate for some projects, traditional "waterfall" development cycles are increasingly regarded as too rigid and time-consuming, making it harder for organizations to accelerate use of analytics and respond to changes in requirements during development cycles.

Waterfall development cycles are made up of a sequence of steps for requirements capture, analysis, design, coding, integration, testing, installation, and maintenance. Waterfall methods have been most successful where there are clear, predetermined, stable objectives; in BI, traditional enterprise reporting can fit this description. However, analytics-oriented projects typically have less clarity; data scientists, analysts, and business subject matter experts need to explore data, try different variables and combinations of data sources, and so on. Organizations are finding that agility is an important attribute of data-intensive development projects—and so, not surprisingly, agile methods have become popular.

We asked research participants what methodologies, if any, their organizations are using for BI, analytics, AI, and/or data integration and management projects (Figure 3). Agile methods are clearly the most popular, with 78% of organizations noting that they use them compared to waterfall methods (46%).

Agile methods such as kanban have been part of software development culture for nearly 20 years and build on a long history of software engineering efforts to streamline development, increase collaboration among stakeholders, and improve quality and ROI. About one-quarter (24%) are using lean, a method related to agile that is inspired by the Toyota Production System methodology. About half as many (13%) are using process models such as Capability Maturity Model Integration (CMMI) and ISO 9000 quality management systems standards.

### Is your organization using any of the following methodologies for BI, analytics, AI, and/or data integration and management projects?

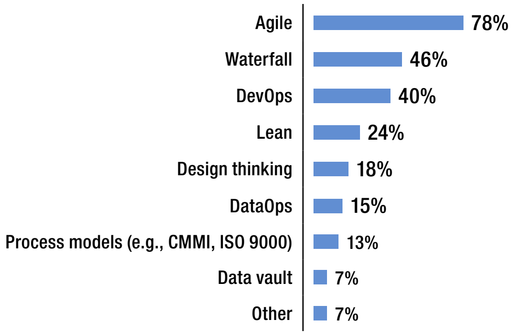


***Figure 3.*** *Based on answers from 136 respondents. Respondents could select all answers that apply.*

A considerable percentage of organizations surveyed use DevOps practices (40%), which extend agile methods to increase automation of repeatable software development tasks and overall enable faster and more reliable releases. DevOps, like agile, aims at facilitating collaboration between stakeholders—in this case, between developers (“Dev”) and operations (“Ops”) personnel such as IT systems engineers, database administrators (DBAs), and network and security personnel. DevOps and agile practices bring stakeholders together to work in self-organizing (rather than strictly IT-led) teams that meet continuously to test iterations and produce deliverables regularly instead of waiting until the end of a waterfall cycle.

**Design thinking is helpful for capturing human factors.** Almost one-fifth (18%) are using design thinking, a discipline that builds on methods that have been used by organizations to learn human factors in customer experiences and spur innovation to improve experiences. Organizations are applying design thinking to bring greater fulfillment to internal customers, i.e., users of BI and analytics applications.

Rather than focus on traditional requirements gathering for data access, query needs, and other technical aspects, design thinking helps development teams understand human and emotional factors that often have the biggest effect on the success or failure of applications. Design thinking's five phases—Empathizing, Defining, Ideating, Prototyping, and Testing—can complement agile and DevOps methods.[1]

**DataOps for increasing collaboration and agility.** DataOps, as the term suggests, picks up ideas from DevOps to apply them to how an organization's developers, IT, business, and other stakeholders can improve collaboration throughout data life cycles. DataOps is a relatively recent term; Figure 3 shows that currently just 15% of organizations surveyed are currently using this method. A central problem it addresses is the delays and inefficiency that exist due to poorly integrated development, transformation (ETL and ELT), data quality, and enrichment steps as well as governance and other data life cycle phases.

With organizations needing to integrate these complex and interdependent phases into data pipelines that serve analytics and AI development, eliminating inefficiency and applying automation is critical. DataOps can offer a framework that enables organizations to get the big picture of multiple and diverse data life cycles and set in motion continuous improvement cycles for how they source, profile, transform, and ultimately deliver data for end purposes. DataOps can also facilitate better stakeholder collaboration and teamwork to achieve outcomes. With shared principles and approaches, DataOps can complement use of agile methods and the use of design thinking practices.

## Defining Objectives for Shortening the Path to Value

Only about half of those surveyed say their organizations are currently either good or excellent at delivering data insights that provide business value.

In Figure 4, we can see how survey respondents regard their organizations' success in achieving important project objectives, many of which are addressed by the methodologies and frameworks shown in Figure 3. The ultimate goal of BI and analytics is to deliver data insights that provide business value; however, only about half of respondents regard their organizations as either very (11%) or somewhat (40%) successful at achieving this objective (51% combined). A significant percentage—30%—regard their organization as merely average at achieving this goal and 18% think they are unsuccessful (1% don't know).

We can draw four other important takeaways from Figure 4:

- Only a minority of organizations are successful at collaborating and communicating to share knowledge and feedback, with only 9% calling themselves "very" successful (26% are somewhat successful). This sort of sharing is crucial to learning from mistakes and continuously improving, which 40% of respondents deem their organizations either very or somewhat successful at doing. Agile, DevOps, and DataOps methods emphasize communication and collaboration to support continuous improvement.
- Organizations still have progress to make in setting project objectives, measuring progress, and organizing themselves to achieve deliverables. Only 9% of organizations surveyed are very successful in identifying value measures and quantifiable objectives, while 26% are somewhat successful. Only about one-quarter (23%) are either very or somewhat successful in scheduling clear-cut deliverables from easiest to most difficult. The majority also fall short in defining strategy and following it within alignment with corporate objectives; 38% are successful at this objective.

[1] For more, see *TDWI Checklist Report: Using Design Thinking to Unleash Creativity in BI and Analytics Development*, online at tdwi.org/checklists

How successful is your organization in achieving the following objectives in its BI, analytics, AI, data integration, and data management projects?

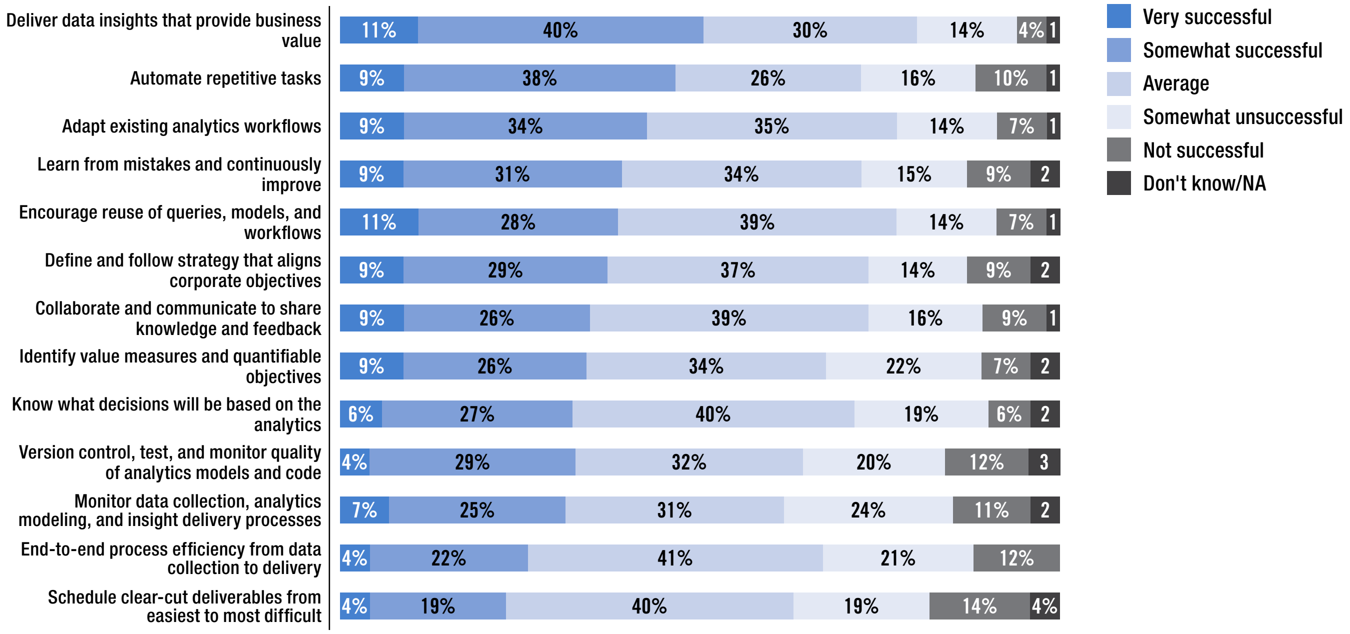


***Figure 4.*** *Based on answers from 140 respondents. Ordered by highest combined "very successful" and "somewhat successful" responses.*

- Almost half of organizations surveyed are successful at automating repetitive tasks (47% combined very and somewhat successful). Fewer are succeeding at encouraging reuse of queries, models, and workflows (39% combined). About one-third successfully monitor data collection, analytics modeling, and insight delivery processes (32% combined). Automation, reuse, and monitoring processes are factors that can lead to better quality and efficiency. Only about one-quarter regard their organizations as successful with end-to-end process efficiency from data collection to delivery (26% combined "very" and "somewhat" successful).
- Adaptability is important in analytics workflows because users may want to add new data sources or make iterative adjustments as they explore data and evaluate models. About one-third of respondents (34%) say their organizations are somewhat successful with adaptability and just 9% are very successful; 35% say they are average.

# Technologies and Cloud Services in Use

Organizations typically use a range of tools and data management platforms throughout data and analytics project life cycles. The range is widening as organizations augment on-premises technologies with cloud-based services or begin migrations to replace those systems in the cloud. Often, the challenges of integrating the technologies will hold back organizations from achieving faster analytics from faster data. Data interaction and integration technologies that were built to fit the constraints of older servers and processing platforms and to serve older types of use cases such as enterprise reporting are not always capable of meeting emerging demands.

To get a sense of the state of technology implementation, we asked research participants how reliant their organizations' decision makers are currently on various tools, platforms, and cloud-based services. After spreadsheets, which are always the most commonly used tools, participants are most reliant on data warehouses on premises; 33% are very reliant and 31% are somewhat reliant on them (figure not shown).

Organizations are using a variety of data integration and transformation technologies to meet equally varied data timeliness and delivery needs.

About the same percentages rely on supporting ETL, ELT, change data capture (CDC), or data integration systems. About half (51%) rely on data pipelines and data preparation tools. Slightly more than a quarter (28%) say their organizations' decision makers rely on data virtualization layers or virtual databases. These technologies enable organizations to access data from multiple sources without the delays involved in physically moving and consolidating the data. Data virtualization solutions do not store data; they federate queries for execution at the sources. In contrast to ETL and CDC technologies, which require time for data movement, transformation, and determination of the changed data, virtualization offers real-time data access.

A significant percentage of organizations (43%) rely on orchestration or scheduling tools, which is an impressive number given the newness of the idea of orchestration to data management contexts. In a musical score, a composer will orchestrate the parts so the musicians' playing fits into the larger concept of the composition. In the realm of data, orchestration aims to reduce confusion by focusing on how fragmented activities fit into the whole: that is, how activities involving the data fit into an end-to-end process from data identification and collection to its use in decisions and actions.

Data orchestration also highlights where automation could improve efficiency and effectiveness in those activities and organize their scheduling to fit business objectives. This is critical as organizations seek to reduce latency and want to get value from near or true real-time data sources.

Our research sees reliance on cloud-based data warehouses as currently less common, perhaps showing the relative immaturity of this option and reflecting the challenges of migrating to the cloud. Just 12% of respondents say their decision makers are very reliant on cloud-based data warehouses and 23% are somewhat reliant on them. Similar percentages of respondents say their organizations' decision makers are either very reliant (11%) or somewhat reliant (27%) on cloud-based BI, online analytical processing (OLAP), or analytics services (provided via software-as-a-service, i.e., SaaS).

Both IT-centric and departmental BI and analytics remain alive and healthy in organizations surveyed. Almost two-thirds of participants (63%) say their decision makers are either very or somewhat reliant on BI reporting and/or an OLAP system managed by central IT. Over half (54%) say they rely on departmentally managed versions of these systems. About half (49%) of organizations surveyed say users rely on embedded dashboards or analytics in business applications, such as CRM, SFA, or ERP.

It is likely that this variety of IT-centric, departmental, and embedded BI and analytics tools will continue to be part of organizations' data interaction environments, even as SaaS and other cloud services are adopted. Organizations should adopt data architectures that balance centralization requirements for governance and data quality with demand for user and departmental control of self-service capabilities.

**More organizations are using or plan to use data catalogs, glossaries, and metadata repositories.** In a 2016 TDWI Best Practices Report survey, only 17% of research participants said their organizations used these systems, which can provide efficient and often better centralized ways of gathering metadata, documenting data definitions, and providing other descriptive, location, and origination information about the data.[2] In the survey for this report, we find that about double the percentage of organizations surveyed (36%) are using these systems.

A slightly larger percentage (39%) are using master data management (MDM), which is both a process and technology system for gathering definitions and knowledge about data resources that are related to higher-level entities such as customers or products. If they are up-to-date, accurate, and comprehensive, data catalogs, glossaries, metadata repositories, and MDM can speed insight by making it easier for all types of users to more easily find related data, integrate it, and analyze it.

**One-third are using real-time alerting and analytics.** Alerting and notification are valuable, particularly in operational use cases where managers and frontline personnel monitoring a process need to know immediately about changes in the data or when situations arise that demand immediate attention. Alerts and notifications may be delivered in the context of metrics, key performance indicators, predictive analytics, or as embedded functionality in business applications and processes to make it easier for personnel to determine what action to take. One of the key challenges with alerts and notifications is making sure they are important, timely, and relevant; otherwise, personnel can suffer "alert fatigue" and stop paying attention to them because it is unclear whether or why they matter.

Real-time alerting and analytics are valuable for operational use cases; just over one-third of organizations surveyed say decision makers rely on these technologies.

Real-time analytics, which will be discussed later in this report, can be critical to developing smarter alerts and notifications so that "faster" does not just result in too much information for personnel. An objective of real-time analytics is to analyze data as it is received to find significant patterns, anomalies, and trends. Alerts and notifications can pick up these insights and inform personnel who need to know and are accountable for taking action. Analytics insights should be presented within the context of the receiver's responsibilities. Our research finds that just over one-third (36%) of research participants say decision makers in their organizations rely on real-time alerting and/or analytics, although only 6% are "very" reliant.

## Delivering "Personas" the Right Data at the Right Time

One of the most challenging aspects of creating a data architecture is balancing the needs of different types of users with broader enterprise data management and governance requirements. Self-service technologies have enabled users to personalize their data interactions so they are no longer limited to monolithic, one-size-fits-all enterprise reporting. However, an imbalance toward too much self-service can lead to a plethora of data silos, too much data and ETL redundancy, and governance problems. This imbalance can thwart progress toward reduced time to insight.

To find the right balance, it is helpful to identify *personas*: the roles, data and analytics needs, common likes and dislikes, data access and sharing authorizations, and daily decision-making challenges of different types of users. Then, development teams can have a better understanding of the requirements that different types of users have. Organizations can bring stakeholders together in a center of excellence (CoE) or governance committee to make sure that the defined personas are accurate, assess whether new definitions are needed (such as for external partners and customers), and plan how to make improvements to the personas' data and analytics experiences.

We asked research participants about the satisfaction levels of different types of personnel with their ability to access the data and information they need, when they need it, for analytics, visualization, or other data consumption (Figure 5). Business and data analysts are the most satisfied, with 11% of respondents saying these personas are very satisfied and 55% saying they are somewhat satisfied.

[2] *TDWI Best Practices Report: Improving Data Preparation for Business Analytics*, Q3 2016, page 10, online at tdwi.org/bpreports

Thinking of your organization’s most recent projects, how satisfied are the following types of personnel with their ability to access the data and information they need when they need it for analytics, visualization, or other data consumption?

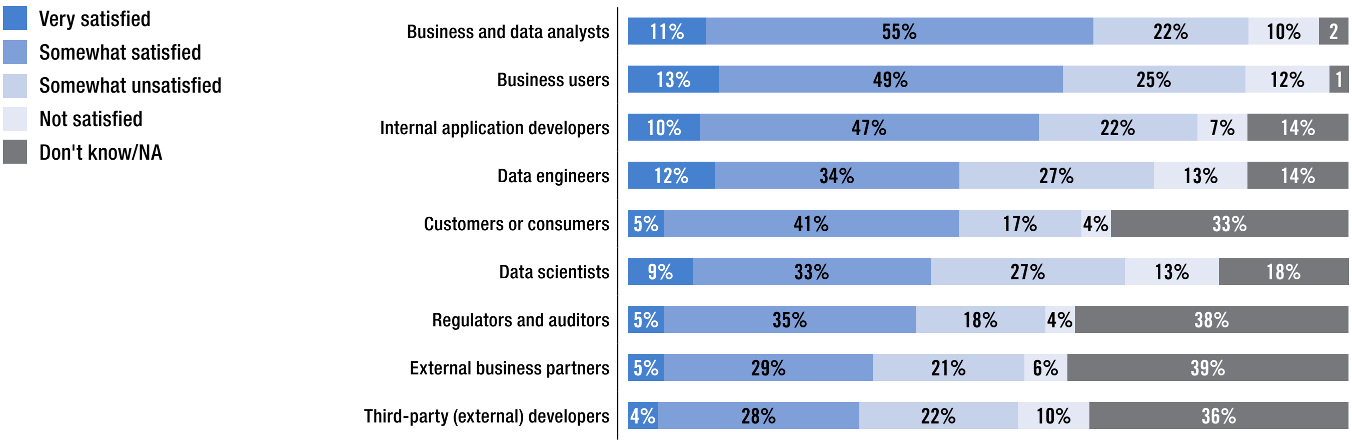


***Figure 5.*** *Based on answers from 136 respondents. Ordered by highest combined “very satisfied” and “somewhat satisfied” responses.*

These personas, of course, can be highly varied. Some business and data analysts are experienced in using BI and analytics tools and working closely with the data while others are not technical and do not have the time or interest to acquire programming, data integration, transformation, and other skills. Typically, they are responsible for providing trusted and relevant subject matter data, visualizations, and analytics to business executives and managers. In other words, easier, more automated access to trusted data is essential for business and data analyst personas.

Data scientists at organizations surveyed are less satisfied than business and data analysts (42% combined “very” and “somewhat” satisfied). They also need tools and practices that enable them to prepare data faster so they can focus on analytics, machine learning development, and providing insights to business leaders. However, along with ease of use they need flexible access to many types and sources of data as well as readily available compute and processing power to test analytics models and algorithms.

Research participants surveyed indicate that business users in their organizations are also reasonably satisfied, with 13% very satisfied and 49% somewhat satisfied. Business users are generally data and analytics consumers, not developers; ease of use and data relevance, including timeliness, are critical. Data engineers are at the other end of the spectrum; they possess technical skills and knowledge about the data. A little under half of research participants say data engineers in their firms are satisfied (46% combined “very” and “somewhat” satisfied). They require technologies and practices for preparing, integrating, and processing data faster, orchestrating and operationalizing BI and analytics, and improving data quality.

Self-service and embedded analytics continue to be high priorities as organizations seek to empower decision makers with relevant and timely data insights.

## Importance of Self-Service and Embedded Analytics

TDWI research finds that organizations continue to regard achieving higher levels of user self-reliance as a very high priority; 53% of research participants say it is very important and 36% say it is somewhat important. (No figure shown.) Self-service capabilities enable analysts and users closest to business questions that need answering to shape their data interaction and visualization. Modern self-service technologies use automation, AI-driven recommendations, notifications, and intuitive interfaces to relieve decision makers of either knowing the intricacies of how to set up data interaction themselves or going to IT developers for every need. They can then move faster to develop relevant data insights.

Self-service analytics and visualization are becoming more mainstream in embedded reporting and data interaction functionality. Embedded (sometimes called "inline") BI and analytics functionality is critical for users who lack the skills, time, and interest to work with dedicated tools and are more comfortable staying within their business application such as CRM, SFA, ERP, a mobile app, or specialized, vertical industry software solution.

Self-service technologies now enable not just third-party solution developers but also organizations' developers themselves to embed analytics in cloud-based services for customers and business partners as part of strategies to monetize data and analytics. Organizations may also develop such services for internal employees, sometimes supplementing access to internal data with access to external syndicated data about customers, suppliers, or other subjects of interest.

We asked research participants if their organizations currently embed analytics into any of the applications or systems listed in Figure 6. We can see that the largest percentage embeds analytics in dashboards or reports (72%), followed by performance management KPIs or scorecards (51%).

### Does your organization currently embed analytics into any of the following types of applications or systems?

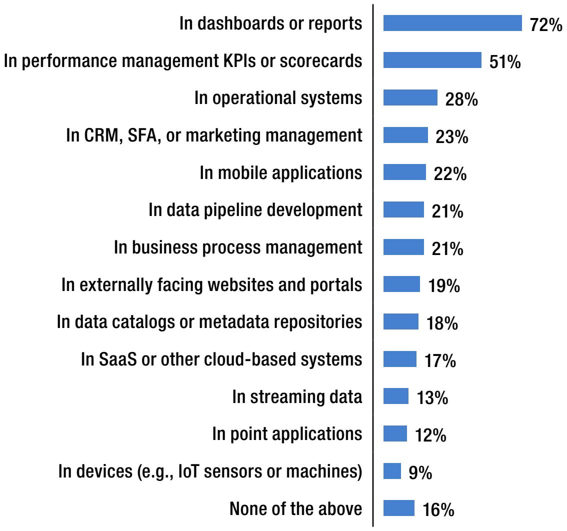


***Figure 6.*** *Based on answers from 134 respondents. Respondents could select all answers that apply.*

Tightening integration between analytics and metrics can enable those accountable for the metrics to ask questions about the data and look at trends, patterns, and predictive insights to determine the right course of action sooner. The research shows that some organizations are embedding analytics for externally facing websites and portals (19%), potentially as part of data monetization strategies. Leading-edge organizations are beginning to embed analytics in collaborative, workflow, and instant messaging systems such as Slack.

In addition, although not the majority of research participants, some organizations surveyed are also embedding analytics in data pipeline development (21%) and data catalogs or metadata repositories (18%). These uses of analytics, as well as AI and machine learning, can help organizations profile data as it moves from sources through pipeline steps for transformation, enrichment, and delivery to users. They can use analytics to look for patterns and anomalies in raw source data, learn the data's quality, and in general speed up preparation steps. AI and machine learning enable organizations to deal with higher volume, velocity, and variety of data as well.

**Some embed analytics in operational, streaming data, and Internet of Things (IoT) systems.** Just over one-quarter (28%) embed analytics in operational systems; 21% embed analytics in business process management. Along with potentially reducing the latency between analytics and action, embedding analytics in these systems enables users to view data insights within the context of their operations and processes. Enabling users to interact with data in context is important to increasing an organization's overall data literacy.

Embedding analytics can become necessary for organizations that want immediate notifications or predictive insights from streaming data, such as IoT data, for operational monitoring and management. Whether fully embedded or not, having analytics more deeply integrated with applications or business process management systems can reduce decision latency; personnel do not have to move to a different tool or interface to consume the analytics. Just 13% of participants are currently embedding analytics in streaming data and 9% are doing so in devices such as IoT sensors or machines.

## User Satisfaction with Data Interaction

Whether they are using cloud-based or on-premises self-service tools or embedded functionality, the key question is: Are users satisfied with how well they can perform data interaction activities with their BI and analytics tools or cloud-based services? We asked research participants this question regarding a range of activities (see Figure 7). The largest percentage indicated reasonable satisfaction with visual analysis of data, with 15% very satisfied and 48% somewhat satisfied.

Research indicates reasonably strong satisfaction with capabilities for visual analysis of data; less so with data integration, blending, and preparation.

How satisfied are users in your organization with how well they can perform the following activities with their software tool or cloud-based service(s)?

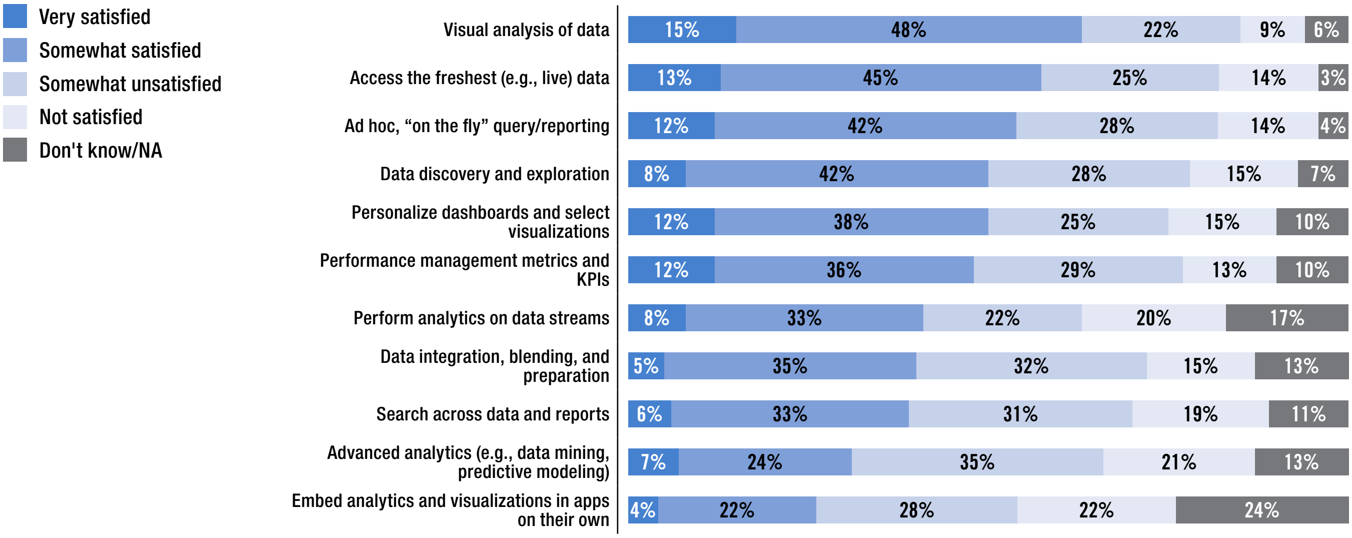


***Figure 7.*** *Based on answers from 132 respondents. Ordered by highest combined "very satisfied" and "somewhat satisfied" responses.*

Performing visual analysis goes beyond just consumption of visual presentation of data; depending on the use case, it could require functionality for filtering, drawing comparisons, examining correlations, visual pivoting and summarization, and drill-down data exploration. Half of research participants (50%) are either very or somewhat satisfied with their data discovery and exploration. This suggests that users are either bumping up against the functionality limits of their tools or may need additional training to make full use of the functionality at their disposal.

It could also mean that many users are not happy with the data sets available to them. Organizations surveyed indicate that there is significant room for improvement regarding data integration, blending, and preparation. Only 5% are very satisfied and 35% are somewhat satisfied; 47% are unsatisfied, with 15% answering "not satisfied."

**Moderate satisfaction seen for access to freshest data and data streams.** More than half of participants say users in their organizations are either very or somewhat satisfied with their access to the freshest (e.g., live) data (58% combined). Access to this data is important to users in finance, risk management, and operations, for example, who need to monitor current business transactions and potentially fraudulent events and receive updates about situations in near real time. Rather than wait for users to query the data, some systems can push notifications to a dashboard on either a desktop or mobile device. When notified, typically users will want to compare the latest data with historical data in a data warehouse; some may want to perform deeper analytics across sources to answer business questions prompted by the notifications.

Organizations that need more than access to live data and notification systems are advancing beyond standard tools to position analytics directly on real-time data streams. Technologies built with open source frameworks such as Apache Spark, Kafka, and Flink are enabling more organizations to attempt such projects. Streaming data sets can be sourced from event-streaming platforms such as Apache Kafka, for example, or directly from sources such as IoT sensors, equipment monitors, pricing systems, fraud detection systems, and online customer behavior.

Although some platforms enable access to real-time data through traditional Open Data Connectivity (ODBC) and Java Database Connectivity (JDBC) APIs to BI dashboards, many organizations find they need to augment data management with technologies that are specialized for managing real-time data streams. In Figure 7, we can see that although not the majority, a significant number of organizations are succeeding with analytics on real-time data streams; 41% of research participants say users are satisfied with their ability to do this, although of those only 8% are very satisfied.

**Most users are satisfied with ad hoc query and reporting but less so with advanced analytics.** Just over half of research participants say users in their organizations are satisfied with ad hoc, "on the fly" query and reporting (12% very satisfied plus 42% somewhat satisfied). BI and analytics tools typically offer easy-to-use capabilities that enable users to deal with dynamic needs on their own, often by choosing visualizations, filtering data, and other personalized attributes; some tools enable on-demand data integration and transformation. Although the intention of ad hoc functionality is to meet immediate needs, often ad hoc activities can turn into new requirements for standard reports and analysis. Organizations should monitor ad hoc querying and reporting to determine if user demands could be met more effectively—and developers' talents and system resources could be used more effectively—with standardization.

**Only 31% of research participants say their organizations' users are satisfied with advanced analytics, most likely indicating continuing dependence on highly skilled data scientists.**

Research participants show less user satisfaction with advanced analytics such as data mining and predictive modeling, with 7% very satisfied and 24% somewhat satisfied. Most users are likely still dependent on highly skilled data scientists for analytics beyond what they can do with self-service tools. Data scientists are capable of exploring and blending broader types of data and can build homegrown models and AI and machine learning routines. They can also push beyond limited types of advanced analytics supported by tools and perform their own statistical and mathematical analysis.

Long-term technology trends show incorporation of more advanced capabilities in leading toolsets in the coming years to enable "citizen data scientists"—advanced business and data analysts as well as power users who want to go beyond standard BI and OLAP capabilities—to do more on their own. However, the research results indicate that users are not yet highly satisfied with the advanced analytics functionality at their disposal.

## Overcoming Barriers to Faster Value from Data

Users face obstacles when trying to access the right data at the right time for analytics, AI development, and visual data consumption. We asked research participants what are the biggest challenges their organizations confront when they are trying to use data assets faster and more effectively. As we often discover, the number one challenge is poor data quality and completeness; 67% of research participants cite this as their biggest challenge (figure not shown).

Data quality is the biggest barrier to gaining faster value from data, according to research participants; disconnected data silos is the second-most common barrier cited.

The pace of BI, analytics, and AI projects typically slows down considerably when users and analysts confront too much dirty data and not enough lineage information to understand and fix anomalies. By taking steps to improve data quality and establish data lineage, organizations can enable projects to move faster, with fewer mistakes and disputes about whether to trust the data.

Data quality problems often go hand in hand with the problem of too many disconnected data silos; 51% of research participants cite this as one of their biggest challenges. To fix the silo problem as well as improve data quality and consistency, organizations often will seek to consolidate and integrate selected data into a central physical location such as a data warehouse, which today could be on premises or in the cloud. However, as the number and volume of data sources and silos grow, rather than try to consolidate all data, many organizations will create a data architecture that combines consolidation with a greater role for data virtualization and data catalogs. A virtualization layer provides an alternative to relying on heavy data movement and consolidation by enabling federated querying and virtual, single views of multisource data. Centralized data catalogs bring together metadata to make it easier to find data and keep track of its lineage.

Almost half of research participants (46%) say data governance and regulatory concerns are one or their biggest challenges in enabling data assets to be used effectively. In interview research TDWI finds that analytics and AI projects often run into barriers when IT lacks confidence that it can govern how the data will be used and shared. For users, governance is an important element of trust; if users are uncertain about how they can use the data or whether they can gain regular access to it, they will not trust it. Governance and data privacy regulatory adherence should be priorities at every stage of a project's evolution.

**Problems exist with data pipelines, transformation, and preprocessing.** About a third of respondents say data transformations that are slow and difficult to manage can impede their realization of value from data (34%). Traditionally, ETL development has involved significant manual coding and has been time-consuming for IT to manage as the number and complexity of ETL routines increase. ETL routines often slow down or have to be restarted when data formatting and other data consistency issues across sources cannot be resolved without human intervention. Modern solutions are offering greater automation, which can help organizations streamline transformation and alleviate some manual work.

Organizations also have the option of using data virtualization alongside ETL processes to provide faster exploration of new data. They can then transform the data as needed rather than as a condition of it being loaded into a target database such as a data warehouse.

As organizations move beyond classic BI and OLAP requirements to address new, analytics-oriented transformation needs and bigger data volumes, our research shows that organizations are having difficulties with data pipelines and transformation. More than a quarter of research participants (28%) regard difficulties in updating, maintaining, or iterating data pipelines as one of their big challenges. It appears that organizations are faring somewhat better with data preprocessing, such as for OLAP and reporting; only 17% of research participants say slowness and lack of scalability for data preprocessing is a challenge.

**APIs and pipelines offer an alternative to traditional data integration.** The growing popularity of application programming interfaces (APIs) is also putting pressure on traditional ETL and electronic data interchange (EDI) systems, often forming a reason why organizations choose to adopt data pipelines. APIs enable applications to expose data that the application is allowing to be shared with other applications. APIs make it easy to establish data connections, request data, review documentation about its structure, and then with permission, automatically populate an application with data from another application. Data pipelines can build data security, preparation, transformation, versioning, deduplication, and other required activities into the interchange of data through APIs.

APIs, along with data pipelines, can create simpler connections that allow data to flow easily between applications. Their use does not require the knowledge about data sources needed for traditional ETL processes, including knowledge about any changes in the data sources and structures that might require ETL processes to be rewritten. The API/pipeline style fits with the cloud computing paradigm, which favors easier and simpler data connectivity. However, this ease and simplicity depends on adoption of open, standard APIs where possible; otherwise, APIs can begin to resemble old-fashioned, point-to-point integration that requires specialized knowledge about each API. Big companies as well as industries such as fintech are working to establish open APIs.

**Lack of skilled personnel and investment are significant obstacles.** Over half of research participants say not having enough skilled personnel is one of their organizations' biggest issues (55%). A shortage of skilled personnel is often a strong motivator for organizations to look for solutions that can automate steps in making data assets valuable. It is also a driver behind the trend toward cloud-based services for data integration and management that can obviate the need for adding skilled personnel onsite who can work with on-premises systems. A significant percentage also say overall investment is inadequate to the challenges they face (41%).

## Provisioning New and Fresh Data and Updating Analytics Models

Although some data sets are constant and stable, others are always changing. In addition, for analytics and AI development, analysts, data scientists, and other users typically want to examine new sources alongside current ones to evaluate correlations and look at different variables. One-fifth of research participants (20%) say the problem of data being too old and not refreshed often enough for these purposes is an obstacle to getting value from data.

We asked research participants how long it takes on average to add new data to their data warehouse to be available for reporting, dashboards, and analytics (figure not shown). Some could do so relatively quickly; nearly one-fifth (18%) say it takes between one and two weeks on average.

**The largest percentage of respondents say it takes between one and three months to add new data to their data warehouse; 18% can do this in one to two weeks.**

When we asked this question in 2012, 15% said new data could be added within this time frame, so the percentage of organizations that can add new data at a fairly fast rate has increased slightly.

In this current study, the largest share of respondents (28%) say it takes between one and three months to add new data. In 2012, a slightly larger percentage (31%) said it took this amount of time; the results again indicate that organizations are getting a little quicker about adding new data to their data warehouses. Data warehouse automation tools have become more prevalent in use since 2012, which may be helping organizations add new data and set up and populate tables and columns sooner.

**Addressing changing dashboard and reporting requirements.** In addition to requests for new data, over time users have different needs and preferences for how they interact with data through dashboards and other data consumption applications on desktops and mobile devices. In the best case, developers have interpreted what users want and have produced applications that are stable and allow for incremental adjustments, such as to visualization styles. However, users' requirements often change more substantially. If dashboards, reporting, and other applications are no longer in sync with users' data interaction needs, they will stop using them.

Our research indicates that dashboard and reporting requirements are fairly stable. About half of research participants (49%) say less than a quarter of users' requirements change monthly if not more frequently (figure not shown).

Comparatively, about half as many (25%) say between one-quarter and half of dashboard or reporting requirements change that often. Just 16% of respondents say more than half of users' dashboard or reporting requirements change at least monthly (10% don't know). Our interview research finds that the features most subject to change are how the data is represented visually, options for filtering the data, and adjustments needed to KPIs. Users also want to see their dashboards and reporting upgraded with the latest technologies for drag-and-drop interfaces, search, and pop-up data and visualization recommendations.

Organizations increasingly need to support data demands of AI/machine learning and on-demand analytics. Fortunately, current requirements appear to be fairly stable for these projects.

**Requirements for machine learning and/or on-demand analytics are fairly stable.** Moving beyond dashboards and reporting, the latest set of requirements that many organizations must address are those for machine learning and for business-driven, on-demand analytics. These two areas may come together if AI techniques such as machine learning are embedded in on-demand analytics, which are typically packaged solutions or cloud-native services designed to address specific business requirements such as customer segmentation, sales opportunity analysis, pricing, or a vertical industry need.

We asked research participants what percentage of their organizations' machine learning models and/or on-demand analytics requirements change at least monthly, if not more frequently. The largest percentage (39%) say only one-quarter or fewer of their requirements change that frequently (figure not shown). Just 15% of respondents say between one-quarter and half of requirements for these systems change monthly, if not more frequently, and only 8% said more than half of requirements change that often. However, 38% of respondents either don't know or find the question not applicable, which indicates that machine learning and on-demand analytics are not yet mainstream. As the use of these technologies and solutions grows and they begin to serve a greater variety of use cases, we may see less stability in the rate at which requirements change.

## Preparing, Transforming, and Cataloging Data

In a data-driven world, reducing the latency between when data is created and collected and when it can be used for analytics, notifications, machine learning, and other uses is critical. However, data preparation—that is, the steps that must occur to make raw data ingested from one or multiple sources usable for users and applications requirements—can be notoriously slow. Data profiling, quality improvement, transformation, enrichment, and other steps that are part of data preparation procedures in data pipelines can take up the majority of users' time, not to mention IT specialists who work on bigger jobs for preparing data and building pipelines for data scientists, analysts, and other users.

Thus, it is not surprising that organizations surveyed show high interest in making improvements. The majority of research participants (74%) say it is either "extremely" or "very" important for their organizations to reduce the amount of time and resources spent on data preparation, transformation, and pipeline processes (figure not shown).

We asked research participants what percentage of the total time spent on recent BI and analytics projects was devoted to preparing the data compared to the time spent performing analysis and data interaction (see Figure 8). The survey results are fairly consistent with what we saw in 2016 when we asked this question. In this report, however, research participants indicate that an even higher percentage of users' time is spent preparing the data than it was in 2016. Almost half of respondents (49%) say 61% or more of users' time is spent on data preparation; in 2016, 45% said this amount was being spent. Thus, it appears that organizations may be losing rather than gaining ground on objectives for reducing the time spent on preparation, transformation, and pipelines.

### Thinking of your organization's most recent BI and analytics projects, what percentage of the total time was spent preparing the data compared to the time spent performing analysis and data interaction?

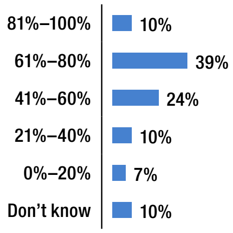


***Figure 8.*** *Based on answers from 130 respondents.*

Manual data preparation and transformation on spreadsheets continues to be a big contributor to slow and mistake-filled BI and analytics projects. TDWI finds that many organizations still cleanse and enrich data on spreadsheets. This can present consistency and quality problems because typically little standardization and documentation exists for the macros, formulas, and processes undertaken to prepare the data. In addition, as organizations begin to use new data sources such as streams of web logs, IoT sensor data, mobile device data, or geolocation data, they are running into difficulties capturing and interpreting it using spreadsheets. Given the volume and speed of this data, they may not be able to capture, for example, the specific points in time in continuous data streams that are needed for time-series analysis.

If errors and inconsistencies become embedded in the data used for analytics and visual reporting, it can take weeks or months for organizations to uncover them, which can have many negative repercussions, not the least of which is a lack of confidence in the data. It is important for organizations to invest in solutions that can handle data preparation, including diverse transformation requirements, especially as they begin to work with fast, continuous, and high-volume streams of data for both alerting and analytics. Such solutions can reduce bottlenecks and streamline the flow of data from ingestion through preparation to support a variety of use cases.

## Variety of Options for Data Pipelines

Modern data pipelines use automation and AI and are faster, more efficient, and capable of addressing quality and transformation needs.

Data pipelines are about creating a flow of data from the original sources through different stages in preparation and transformation for its ultimate use by an organization's variety of users, analytics, AI programs, and business applications. Organizations also develop pipelines to populate cloud-based data warehouses, data lakes, and storage or to move on-premises data into the cloud.

Modern pipelines use automated tools to make the data flow through them faster and more efficiently. Automation and AI can help organizations manage and run large numbers of complex, high-volume data pipelines that have numerous dependencies. Organizations need monitoring tools to inform data engineers and administrators about how a change made to a data structure or a transformation used by one pipeline might affect other pipelines. Organizations also need options in their pipeline management tools for whether to run pipeline processes in regular batch mode, continuous microbatch loading, or real-time data streams.

Many data pipelines perform transformations, which makes them similar to ETL processes. However, a data pipeline is considered to be a broader idea than traditional ETL because pipelines can involve a greater variety of data, including data streams, and the destinations for the data can vary from AI and machine learning algorithms to predictive analytics models, standard BI dashboards, and business applications.

Running largely in slower, traditional batch modes, ETL focuses on extraction of structured data from known sources to a staging area (such as a specialized ETL server) for transformation and then loading into a data warehouse or data mart. For pipelines that have a data warehouse, data mart, or BI/OLAP system as their destination, transformations will be the expected centerpiece. Organizations may choose to use change data capture (CDC) technology instead of ETL to reduce delays because ETL processes are often time-consuming.

Another option is to alter the ETL sequence to extract, load, and transform (ELT). Organizations will use ELT if the target destination (such as a data lake, data warehouse, or analytics appliance) has a powerful data engine that uses massively parallel processing (MPP) and/or clustering and can perform "push down" transformations to be processed in the database rather than in an intermediate station specialized for ETL.

The choice of ETL or ELT often depends on the amount of data and the complexity of transformations. Some modern data transformation and pipeline systems are able to automatically determine which approach is optimal so administrators do not have to manually decide for each workload. ELT is typically the choice for organizations that need faster ingestion for analytics and machine learning programs. ELT processes can land the data, and from there analytics and AI programs may apply custom transformations as part of processes for exploring the data for patterns, trends, and other insights.

Data virtualization is an additional option. Data virtualization leaves the data in place and does not require moving the data from sources to ETL staging areas and then to data warehouses or other target zones. Data virtualization can offer greater agility than traditional ETL because it does not force data coming from multiple sources to comply with a single data model and transformation plan. The data virtualization layer provides single views of logically integrated, multisourced data and supports federated queries to those sources. Virtualization layers can optimize query processing at the sources to take advantage of local processing power.

Data virtualization, ETL, and ELT are not mutually exclusive choices. Organizations will often use a combination of these options in their preparation and pipeline processes to fit different use cases, levels of real-time data requirements, data volumes, transformation complexity, and more. In addition, organizations should evaluate large-scale data processing tools and frameworks such as Apache Spark to scale beyond ELT. Technology solutions are available that can help organizations overcome skills gaps that may make them hesitant to try Apache Spark frameworks and libraries as their data processing needs outgrow traditional technologies.

**Data virtualization, ETL, and ELT are not mutually exclusive choices; organizations will often use a combination of these options.**

**Expense and scalability stand out as challenges.** We asked research participants which of six issues that TDWI frequently sees as major challenges regarding ETL, ELT, and data pipelines are the most important for their organizations to address (see Figure 9). Participants indicate that all are important, but at the top of the list are expense and scalability, with 88% and 87% of participants respectively saying these are combined "very" or "somewhat" important issues.

The potential to reduce expenses is, of course, a driving reason why many organizations want to migrate all types of systems and applications to the cloud, including those for data integration and transformation. Cloud platforms can relieve organizations of having to dedicate budget and resources to storing and processing data on premises, but often they must still develop and execute data preparation routines and pipelines. Utilizing software automation along with cloud options can help reduce expenses generated by intensive manual work by developers and administrators and enable them to focus on more value-adding activities.

### In general, how important is it to your organization to address the following challenges regarding ETL, ELT, and data pipelines?

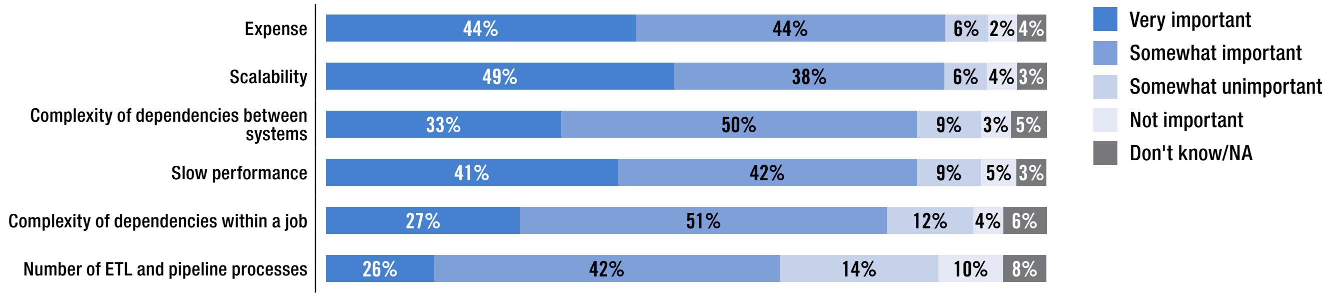


*Figure 9. Based on answers from 129 respondents. Ordered by highest combined "very important" and "somewhat important" responses.*

Software solutions and cloud-based services can help organizations address scalability challenges with ETL, ELT, and data pipelines. Leading solutions today use AI techniques such as machine learning to scale data exploration and analysis of high-volume, high-velocity, and highly varied big data. Slow performance, which 83% of participants said was an important challenge, is often related to scalability issues, although it could be due to other factors such as mistakes and quality errors in the data, queries, and programs. Having several technology options is important for addressing scalability challenges. Organizations can then match the right on-premises system or cloud-native service with the right workload instead of trying to funnel all workloads through one approach such as a traditional data warehouse and ETL process.

Complexity is also a significant challenge. Figure 9 shows high percentages of participants indicating that two aspects of complexity are issues:

- **Complexity of dependencies between systems**. One-third of participants (33%) say this is a very important challenge and 50% say it is somewhat important. If parts of a job fail, such as the loading of a dimension, dependencies between jobs or systems can lead to errors and missing data in fact tables. Administrators often have to manually unload the bad data and restart the ETL or data pipeline process, which can add expense and delays. Job monitoring tools and other automation can help organizations address this challenge.
- **Complexity of dependencies within a job.** Just over one-quarter of participants (27%) say this is a very important challenge and about half (51%) say it is somewhat important. Similar to intersystem dependencies, one of the major challenges in scheduling ETL and data pipelines is accounting for dependencies within jobs. If one does not finish before a dependent job is to begin, it can stall the entire process or could also result in either erroneous or missing data.

The complexity of ETL dependencies and the sheer number of ETL and pipeline processes are major challenges that organizations are facing.

The sheer number of ETL and pipeline processes can also be a challenge; 26% of research participants say it is a very important challenge and 42% say it is somewhat important. As data democratization spreads and leads to more, often concurrent, ETL and data pipeline jobs, organizations need to monitor them for performance and whether they satisfy users. They also need to filter out jobs for reports, dashboards, or other use cases that are no longer important and are taking up resources that could be better used for other jobs.

## Challenges of Moving and Migrating Data to the Cloud

Organizations are motivated to shift more of their analytics, data integration, and data management to the cloud for a combination of reasons. One major reason is so they can move faster to spin up services and cloud-based data platforms to meet immediate business needs. Reducing and controlling costs is another major reason: not just to make analytics and data management less expensive but to make cost issues less of a barrier to data-driven innovation. Organizations also need to scale to handle bigger data volumes, variety, and velocity, as well as the computational power to support more advanced analytics and AI than they can with on-premises systems.

However, to innovate with data and analytics in the cloud and take advantage of the other benefits, most organizations need to move and migrate data to the cloud. These phases can be slow and costly, which can impact how quickly organizations can respond to business needs. Organizations must therefore focus on improving how well they can load, move, and replicate from on-premises sources to cloud-based platforms, potentially use virtualization to reduce data movement and provide views of data in place, and perform operations such as transformations and updates for data used in visualizations, analytics, and AI.

In addition, for most organizations, "the cloud" does not consist of just one platform; our research finds that many organizations use the services of multiple cloud data providers for data storage, data lakes, and data warehouses. The danger is that each one can become a silo, which only exacerbates existing disparate data problems and leads to slower and less complete data access. Thus, as organizations set up data platforms with multiple cloud providers, they should examine what new technologies they will need to move, replicate, and otherwise manage data across them to produce integrated views and access for users.

In Figure 10, we can see the levels of satisfaction research participants have with how well their organizations can accomplish a variety of factors having to do with getting data ready in the cloud for users, analytics, and applications. There is room for improvement; for not one of these factors do we see high "very satisfied" percentages or responses.

### How satisfied is your organization with the following factors regarding loading, replicating, transforming, and updating data from data sources such as on-premises systems to cloud-based data platforms, such as a data warehouse?

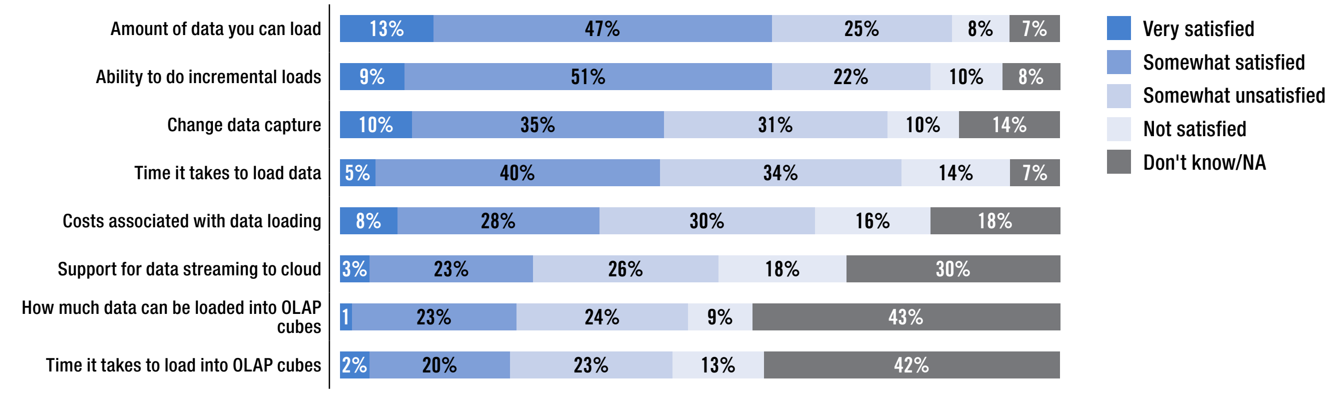


***Figure 10.*** *Based on answers from 131 respondents. Ordered by highest combined "very satisfied" and "somewhat satisfied" responses.*

Organizations show the most satisfaction with their ability to handle the amount of data they need to load (60% combined "very" and "somewhat" satisfied) and their ability to handle incremental loads (also 60% combined). This suggests that there is reasonable but not overwhelming satisfaction with how well organizations surveyed can manage full and incremental loads to cloud data warehouses or data marts and analytics sandboxes. We can also see in the chart that among organizations that indicate that they are loading data into cloud-based OLAP cubes, satisfaction is average regarding the time it takes them to load data into the cubes and how much data they can load into them.

As we saw earlier with cost being the biggest ETL and data pipeline challenge, we find in Figure 10 that the second-highest percentage of dissatisfaction is with costs associated with data loading; 46% are either somewhat unsatisfied or not satisfied. The highest level of overall dissatisfaction is for the time it takes to load data to the cloud (48%), and the third-highest dissatisfaction levels are with support for data streaming to the cloud, with 26% somewhat unsatisfied and 18% not satisfied (note that 30% either don't know or find this attribute not applicable). Instead of loading all data into the cloud in batch, organizations could choose to stream the data into cloud data warehouses or other data platforms continuously. Then, organizations can begin running some analytics or AI programs on this data sooner. However, this research shows that organizations are not yet happy with this option.

Data virtualization can provide an alternative to loading data into a separate, central store from hybrid, multicloud sources. Data virtualization offers transparent access so that users do not need to know how and where to access the data. Data virtualization solutions create a logical view (or "logical data warehouse" as it is sometimes called). Users can then query this view from within their chosen front-end tools. Data virtualization uses metadata extensively, which spotlights the importance of data catalogs and repositories for documenting data knowledge and enabling access to multiple data stores.

## Importance of Data Catalogs and Metadata Management

Metadata—a collection of knowledge about how the data is defined, where it is, its lineage, and how it is related to other pieces of data—has always been critical, but is becoming even more so as data becomes increasingly diverse, distributed across multiple cloud platforms and on-premises systems, and voluminous. Almost all data platforms and sources have metadata; perhaps one of the hardest parts is collecting, maintaining, and updating a complete and accurate centralized view. In recent years, technologies have emerged that provide smart, AI-augmented data catalog and metadata repository development and management. Organizations are becoming better able to give users faster data discovery, self-service access to metadata, and business-driven analysis of data relationships.

The top goal organizations have with data catalogs and metadata management is to make it easier for users to search for and find data, followed by improving governance.

Data catalogs, metadata repositories, business glossaries, and emerging semantic integration can help organizations meet many goals. We asked which goals are most important to organizations. The one selected by most research participants is making it easier for users to search for and find data (79%; figure not shown). This result is consistent with the response to a similar question that we asked in 2018.[3] The goal with the second-highest percentage is also the same as in 2018: improving governance, security, and regulatory adherence (70%). For governance, monitoring access to sensitive data, and data lineage, it is essential to know where the data is, its life cycle in the organization, and how it is being shared.

To adhere to regulations, organizations typically need to audit data management and be able to demonstrate that they are protecting the data. Modern solutions' capabilities for automated tagging and data lineage tracking can be key to making governance effective, easier, and less expensive. Over half of respondents (55%) see the ability to centrally monitor data usage and lineage as a key goal. About one-third of organizations surveyed regard consolidating multiple smaller data catalogs and glossaries as an important goal. Where consolidation is too difficult, slow, and costly, data virtualization integrated with data cataloging can provide an alternative.

A sizeable percentage of research participants are seeking to coordinate data meaning across sources (59%). This is important to resolving debates among users about what data means, whether calculations match up, and—if there are discrepancies among multiple data sources—determining which one is correct. Master data management (MDM) and semantic integration are technologies that build from metadata definitions to bring higher-level concepts into focus and make it easier to find and manage data related to each concept. Coordinating data meaning can ultimately make it faster and easier for users to find and interact with relevant data.

Finally, more than half of organizations surveyed want to use data catalogs and metadata management to improve data preparation and data pipelines (54%). Integration of these technologies plus data governance can help organizations ultimately deliver more trusted and complete data to users, AI programs, and applications.

# Closing in on Real Time: Options for Faster Data

Often the closer users can get to accessing real-time data, the more valuable that data is. Although historical data is acceptable for many use cases, in others (such as personalized customer engagement, fraud detection, and risk awareness) organizations are recognizing that near or true real-time data is the most valuable. A decided majority of organizations surveyed for this report (77%) say near or true real-time data, BI dashboards, and analytics are important to their firm's success, including 30% who say they are very important (figure not shown).

[3] *TDWI Best Practices Report: BI and Analytics in the Age of AI and Big Data*, Q4 2018, online at tdwi.org/bpreports.

TDWI research explored which technologies organizations are using to make data (including data streams) available sooner for BI, analytics, and AI/machine learning (see Figure 11). As when we asked this question in our 2018 Best Practices Report, a data warehouse and/or data mart is again the most common technology used (61%). This is followed by BI/analytics access to live data (48%), which typically means that users can access data as it is being recorded in business applications or transaction processing systems.

**The data warehouse and/or data mart remains the most common technology used to make data available sooner; 26% are using CDC and 23% are using data virtualization.**

### Which technologies are in use by your organization to make data (including data streams) available sooner for BI, analytics, and AI/machine learning?

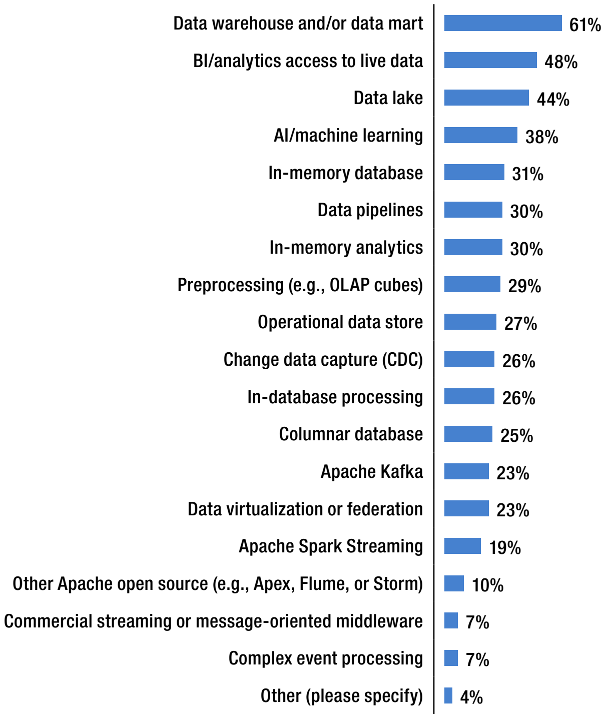


***Figure 11.*** *Based on answers from 124 respondents. Respondents could select all answers that apply.*

To support near-real-time data warehousing, organizations will often use either CDC to update changed data in the warehouse or data virtualization to speed access to sources without having to move or replicate the data to a central physical store. Figure 11 shows that 26% are using CDC, which is about the same percentage as in 2018. More research participants say their organizations are using data virtualization or federation now (23%) than said they were using this technology in 2018 (18%).

Just over a quarter of organizations surveyed (27%) are using an operational data store (ODS)—a system that typically complements a data warehouse by providing users with access to a selected, trusted set of integrated near-real-time data, usually for operational reporting and notification. This is a somewhat smaller percentage than we saw in the 2018 report, which was 36%.

Data lakes garner the third-highest percentage of respondents in this report (44%). Data lakes, increasingly stored in the cloud, are flexible platforms that can contain any type of data. Organizations use them to consolidate data from multiple sources, which could include operational, time-series, and near-real-time data. However, unlike an ODS, data lakes are typically set up for exploratory analytics and AI/machine learning to look for patterns and other insights. Some organizations create operational data lakes or set up portions of their data lake for fast SQL queries (using, for example, SQL-on-Apache Hadoop query technologies) on big data. Organizations can also develop templates and preconfigured views of selected operational data for consistent and repeatable reports or for developing OLAP cubes.

**In-memory analytics and databases are becoming more common.** In 2018, we reported that 21% of organizations we surveyed were using in-memory analytics; this year's report shows that a higher percentage of respondents are using this technology (30%). The use of in-memory database technology has also risen, from 17% to 31% of respondents saying their organizations are using this option.

In-memory platforms, by reducing the need to read data stored on disk, can enable faster access to data for visualization, data exploration, and testing models. As larger random access memory (RAM) caches become available, organizations are able to keep more "hot" data available for computation. Technologies are evolving to make it possible to store entire data warehouses, data marts, or OLAP cubes in-memory. Commercial as well as open source solutions using Apache Spark or the more recent Apache Ignite can support in-memory analytics and database systems. They can also support streaming workloads.

About one-quarter of organizations surveyed are using Apache Kafka, which can support data streaming and real-time analytics and machine learning.

**Organizations are implementing streaming and event processing technologies.** Figure 11 shows that about a quarter of organizations surveyed use Apache Kafka. Now an established platform for distributed streaming of large numbers of events, Kafka began as a messaging system with optimized performance. Organizations are using Kafka to build streaming data pipelines for automated applications that must react to real-time data or must make the data available for analytics and machine learning. Kafka can be a source for Apache Spark Streaming, which 19% of organizations surveyed are using. This module allows organizations to integrate a variety of workloads, including streaming, on the same processing platform, which can reduce programming and modeling complexity.

Only a small percentage say their organizations are using complex event processing (7%), which is an older technology for processing and analyzing real-time event streams. The same percentage of respondents indicate that their organizations are using commercial message-oriented middleware (7%).

Clearly, there is no single technology approach to managing and analyzing near or true real-time data, including data streams. Organizations need to define their requirements for data freshness and the scale of data flows, speed, and volume. Organizations should look at current in-house skill sets and see where they need to hire more experts in data management, data engineering, and data science. They should also evaluate the potential of data management automation and cloud and SaaS options. Organizations should begin with proofs of concept (POCs) and test applications with smaller, well-defined projects.

## Drivers for Faster Data and Analytics

About two-thirds of respondents say actionable information in dashboards would improve if their organizations invested in "fast batch," real-time data, and data streaming.

For project teams to attract investment in technologies and cloud services to enable faster data and support faster analytics, it is essential to articulate carefully the benefits that would accrue to the organization. To learn some of the leading drivers across industries, TDWI asked research participants what would improve if their organizations increased investment in technologies that support "fast batch" and/or delivery of true real-time data, including data streams. About two-thirds of participants (67%) indicate that actionable information in dashboards would improve; more than half (56%) say operational decisions and management would get better (see Figure 12).

### Which of the following would improve if your organization increased investment in technologies that support "fast batch" and/or delivery of true real-time data, including data streams?

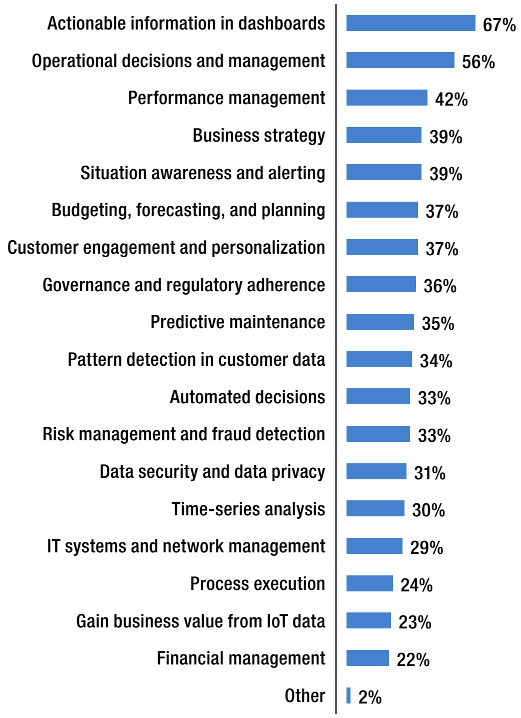


***Figure 12.*** *Based on answers from 126 respondents. Respondents could select all that apply.*

Performance management, frequently one of the main drivers behind dashboard development, is also an area where many organizations surveyed (42%) would expect to see improvement with faster data and analytics. Business strategy, which organizations implementing performance management seek to communicate via dashboards, KPIs, and other metrics, would be a focus for 39% of organizations surveyed.

The results suggest that if and when organizations invest in technologies and services for faster data and analytics, the leading objective is most likely to improve information for managers tasked with increasing operational efficiency and effectiveness. About one-quarter (24%) say process execution would improve with investment in faster data and analytics, which again shows that some organizations see the value of not only faster but smarter business processes. A significant percentage (39%) say situational awareness and alerting are outcomes that their organizations would want to see from investments—objectives that are often key goals behind deployment of streaming data management and real-time analytics.

**Use cases for streaming technologies draw interest.** Streaming data can give organizations new insights into how to solve problems. In Figure 12, we can see that 35% of research participants say their organizations would focus investment in faster data and analytics on improving predictive maintenance; 29% note IT systems and network management would be a priority for improvement. With many organizations now able to tap IoT sensor data from machines and other equipment, they need analytics that can explore this data sooner for trends and patterns that could indicate an imminent failure.

Sensor data is revolutionizing how organizations analyze maintenance problems. Using time-series analysis that combines real-time data exploration with historical data and other information records, organizations can observe changes in sensor data over time. In this way, streaming data can provide organizations with new perspectives on changes over time; our survey shows 30% of organizations want to use faster data to improve time-series analysis.

This modernized analysis can result in smarter maintenance. Rather than use traditional fixed-schedule maintenance, which can either overlook serious problems or apply maintenance when it is not needed, organizations can monitor conditions to see when maintenance is actually needed. They can develop predictive models based on all relevant data rather than assumptions based on a smaller selection of historical data and other records. Similar efficiency could be brought to risk management and fraud detection, which 33% of research participants cite as key areas.

These and other use cases require integrated analysis of real-time, streaming data and historical data. Organizations should evaluate solutions such as data virtualization that can provide combined views of streaming and historical data. Some data virtualization solutions, for example, can read data as it is streaming from edge devices through pipelines for comparison with historical data rather than having to wait for this data to be loaded into a target database.

Predictive maintenance based on IoT sensor data is a growing use case for real-time analytics. Customer engagement and personalization as well as customer behavior pattern detection are other common use cases.

**Faster data drives analytics innovations for business benefits.** Predictive maintenance based on IoT sensor data is a growing use case for analytics in operations, manufacturing, IT, and logistics. However, perhaps an even bigger trend is the use of streaming data to improve customer engagement and personalization, which 37% of research participants indicate is an objective for improvement. Figure 12 also shows that 34% would like to see improvement in pattern detection in customer data. If organizations can analyze near or real-time data, they can respond to customers' interests and concerns in the timeliest manner possible, which is a competitive advantage. Organizations can also gain insights into patterns that they would not find when analyzing only historical data.

## AI for Faster Insights and Automated Decisions

AI techniques, including machine learning and natural language processing, are rapidly becoming part of all kinds of analytics, applications, and data management systems. AI is playing a growing role in enabling organizations to move faster to gain value from data. Until recently, only expert data scientists and developers could use AI techniques. Now many types of users, including consumers, may be using AI embedded in applications and services without knowing it.

AI can help organizations churn through volumes of data to help understand why something is happening, what could happen next, and what to do about it. In some cases, AI techniques are adding the "smarts" to drive fast, automated decisions; in others, AI algorithms are surfacing data insights to augment information humans are using to make decisions.

As we did in 2018, we asked research participants to identify the most important ways in which their organizations currently use (or plan to use) AI such as machine learning to augment BI, analytics, and support data integration and management. The most prevalent choice is to automate discovery of actionable insights (52%; figure not shown); this was also the most common selection in 2018. This could indicate that organizations intend to set up algorithms and models that do not require regular human intervention, but with the end purpose of supplying personnel with insights that can improve daily decisions. Just under half of organizations surveyed (44%) want to use AI to enable faster analytics on large data volumes, which shows that organizations see AI as a solution for scaling up discovery and analytics and growing big data sources.

To augment human decisions, AI-derived insights can be delivered to decision makers in the form of recommendations; 41% of research participants say their organizations want to augment user decision making by giving them recommendations, about the same percentage as in 2018. Some decision makers, however, do not necessarily want recommendations; they just want faster and more comprehensive data search and exploration. Our research finds that 35% of organizations surveyed want AI to help users find, select, and use data for analysis.

To augment human decisions, 41% of organizations surveyed want to see AI-derived recommendations delivered to decision makers.

Because of emerging requirements such as these, organizations must make how data integration solutions use AI a key point in their evaluations. They should examine how AI is applied for faster location, access, and viewing of new data. Some solutions can apply AI programs that learn from and adjust integration and preparation steps to changes in the data and its formats. This can reduce the need for manual adjustments to reconfigure target databases or logical views, which slow down access and analysis. AI programs can also learn from users' search, access, and viewing patterns to recommend related data sets.

**Organizations see a role for AI in data governance, cataloging, and preparation.** Along with helping users locate relevant data and data relationships, research participants see AI helping their organizations better govern, integrate, and manage the data. Survey results show organizations seeking improvements in the following areas:

- **Automating data classification for governance and security.** About a third of research participants (32%) anticipate that AI can help reduce the manual effort and inconsistency that plague data classification and make it hard to locate data for governance and security. Some organizations see AI addressing the general problem of taxonomy development; 20% of research participants regard AI as important for this task.
- **Improve/automate data preparation and enrichment.** A significant percentage of organizations surveyed (31%) currently use or plan to use AI to streamline how data is collected, cleansed, transformed, and enriched for users. Nearly a quarter see AI contributing to the development, management, and maintenance of data pipelines (23%), discussed earlier in this report.
- **Develop and update the data catalog or metadata repository.** Also as noted earlier, collecting and consolidating knowledge about the data and its location and origins is frequently manual and incomplete. Almost a third of research participants (30%) regard AI's role in enabling their organizations to build such a catalog or repository as important; this is up from 27% in 2018.

## Integrating Analytics and Processes; Automating Decisions

In 2018, TDWI research found that enabling greater automation of decisions and actions was a significant driver of new technology investment. Organizations continue to want to identify repeatable decisions that can be programmed into algorithms to reduce delays and costs and increase efficiency and scale. AI does currently and will increasingly play a prominent role in enabling decision automation. However, AI and decision automation are not the whole story when it comes to reducing time to insight. AI is part of a broader trend toward integrating analytics with business applications, processes, and workflows for this purpose.

We asked research participants about the importance of seven different steps that involve integrating analytics to reduce delays and automate decisions (see Figure 13). Integrating visual analytics with business process management topped the list; 77% regard this as either very or somewhat important. With better integration, organizations can bring data insights to bear directly on business processes to improve efficiency and effectiveness. Managers accountable for specific processes can tailor analytics based on their context and knowledge of the data.

### How important to your organization's efforts to reduce delays and automate decisions are the following steps for integrating analytics with business applications, processes, and workflows?

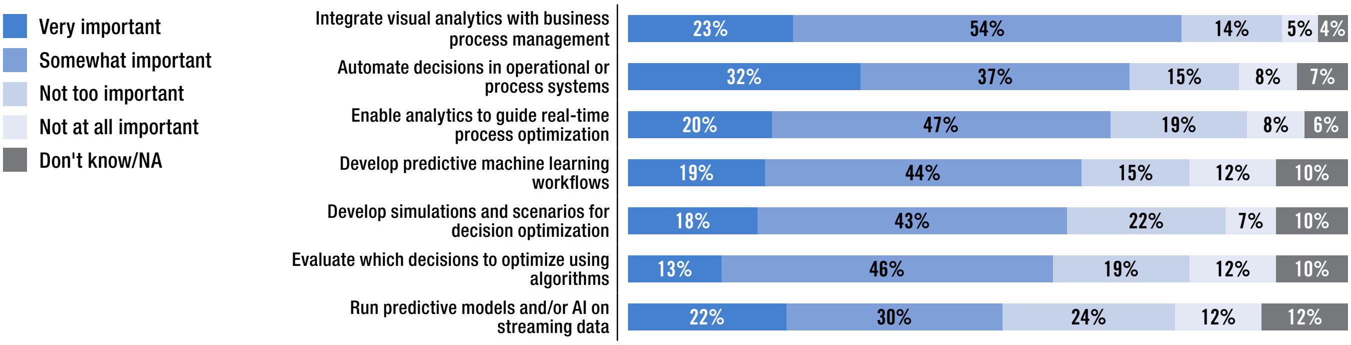


***Figure 13.*** *Based on answers from 125 respondents. Ordered by highest combined "very important" and "somewhat important" responses.*

Demand for tighter integration between analytics and business processes is driving interest in smarter process automation and optimization; large percentages of organizations surveyed see this as a priority.

Demand for tighter integration between analytics and business processes is also driving interest in smarter process automation and optimization. As we can see in Figure 13, more than two-thirds of research participants (69%) say their organizations find it either very or somewhat important to automate decisions in operational or process systems. Nearly the same percentage see it as a priority to enable analytics to guide real-time process optimization (67%).

This level of optimization typically demands continuous data such as data streams to assess the performance of manufacturing systems, for example, and make adjustments automatically. Analytics and AI can help organizations calibrate proper levels based on numerous variables associated with costs, resources, energy used, demand, and other factors. Almost two-thirds want to integrate analytics to enable development of simulations and scenarios for decision optimization (61% combined very and somewhat important responses).

We can see that the activities involving streaming data in Figure 13 are not as commonly regarded as important as those discussed above. However, significant percentages of organizations surveyed do see the steps as important to reducing delays and automating decisions. Just over half regard running predictive models and/or AI on streaming data (52%) as either very or somewhat important. A larger percentage of participants say their organizations want to develop predictive machine learning workflows (63% combined very and somewhat important responses). These workflows could involve many types of data, including real-time streaming data in machine learning development, testing, and operationalization cycles.

# Recommendations

To conclude this report, here are 10 recommendations for developing strategies to reduce delays in data integration and management and increase business value through faster BI and analytics.

Delivering timely data and recommendations to users and finding the right balance between agility and centralized management are key to taking self-service analytics to the next level.

**Deliver more timely data and recommendations to users.** Organizations need to modernize the paradigm for BI, visual analytics, and dashboards, particularly where they are deployed to operational managers and frontline personnel. These users often need very timely data, including in some cases real-time data views and analytics within the context of their responsibilities. They also need applications that are less passive and can supply users with recommendations about data sets that might be relevant, visualizations, analytics, and ultimately prescribed actions to take. Operational managers and frontline users would then be in a better position to make good decisions based on fresher, more contextual, and richer information.

**Find the right balance between agility and centralized management.** An imbalance here leaves neither users nor IT happy and can lead to bottlenecks that thwart faster decision making. Users want agility, and our research shows that organizations are pursuing self-service technologies to give users more freedom in how they personalize workspaces and access and analyze data. However, ungoverned self-service can lead to too many data silos, duplication, and workloads that haphazardly compete for computation and processing. IT's perspective is to ensure good governance, performance, and quality, especially for priority workloads. Yet, clamping down unnecessarily on users will make it harder to move forward and drive users to set up their own data silos in the cloud. Data virtualization solutions could be helpful in reducing dependence on traditional physical data consolidation, which tends to require rigid, preset integration processes. Users and IT need to form committees or a center of excellence to discuss how to balance self-service with centralization.

**Explore how agile, DataOps, and related methods could help projects deliver value sooner.** Too often, organizations are mired in chaotic, inconsistent, and often redundant work in projects for developing BI, analytics, applications, and AI. This can result in delays, inefficient use of data and processing resources, and dissatisfaction among users, who need capabilities as soon as possible. Our research finds that many organizations are using agile (or agile-like) methods—as well as DevOps, DataOps, and design thinking—and are having positive experiences. Organizations that are not using them should try these methods for one or a small number of projects that have clear deliverables to assess whether they are helpful and iron out difficulties before trying them on a larger number of more complex projects.

**Focus on improving data preparation, transformation, and pipelines.** Delays, bottlenecks, and inconsistencies in these areas slow down decision making and force users to spend more time sorting out problems with the data than analyzing it to answer business questions. TDWI finds that there is room for improvement in satisfying user needs in the different phases of data preparation, transformation, and pipeline development. Organizations need to evaluate new technologies, many of which apply AI and analytics to preparation, transformation, and pipeline development to reduce latency and enable users to know more about the data and address problems with it sooner. Advances in tooling combined with use of agile and DataOps methods can improve collaboration and continuous improvement of data so future cycles are more efficient.

**Get the big picture and orchestrate parts into a whole.** As projects grow more numerous and workloads become more complex, it's easy for organizations to get bogged down; then, despite having the latest technologies, it can seem impossible to deliver data faster to support faster analytics. Methods such as DataOps can help organizations get a holistic picture and gain an end-to-end understanding of interrelated steps in projects, stakeholder responsibilities, and where impasses in data flows, updates, and transformations need to be corrected. Organizations need tools that can complement use of DataOps and other methods. Organizations should evaluate tools that not only improve data life cycles but also help them orchestrate what happens in multiple data pipelines and observe the big picture.

**Take advantage of opportunities for automation and reuse.** With an increasing number of analytics and AI workloads needed to meet diverse business demands, it's essential to exploit the potential for smarter automation and reuse in software solutions and cloud services. Organizations heavily dependent on manual coding, monitoring, data preparation, and integration will struggle to scale as more users and applications need to interact with the data and test models and algorithms. "Smarter" is a key word: automating processes will increase the speed and efficiency of routine chores, but organizations should evaluate how AI and analytics can contribute to automating decisions or provide users with recommendations for action.

**Create repositories of knowledge about the data and improve access to them.** Research in this report suggests an upswing in the number of organizations that are developing and managing data catalogs, metadata repositories, and business glossaries, all of which are helpful in bringing together "data about the data" as well as other useful information about its lineage. Such resources are valuable to users, administrators, data scientists, and applications; they can shorten paths to finding and interacting with all data relevant to a subject of interest. Data catalogs and metadata repositories are essential for data governance as well. Technologies are making it easier to develop and use these systems. Organizations should make it a priority to invest in these technologies so they have useful resources of knowledge about their data that users and applications can easily apply to produce more complete views of and access to the data.

**Improve trust in data, analytics, and AI.** It doesn't matter how fast the data, analytics, and algorithms are if no one can trust the data. TDWI finds that a number of issues that impact data trust—most prominently, problems with data quality—are key challenges stalling progress in building strong analytics cultures and accelerating decision processes. Data trust is essential to collaboration on decisions and acceptance of analytics insights. Organizations need to invest in data quality, shared data catalogs, data lineage, and other technologies and practices that give decision makers transparency into the data and confidence in insights drawn from the data.

**Use appropriate technologies to streamline governance.** With goals for faster data and analytics, it's never been more important for organizations to set up rules and policies to protect sensitive data and reduce confusion about where the data is, where it came from, who is accessing it, and what's being done with it. Semantic data integration, which builds on a foundation of central data catalogs and metadata management, can help organizations answer these questions. Organizations should examine options for how a data virtualization layer could help protect access to sensitive data distributed across hybrid, multicloud platforms.

Organizations should develop a strategy for augmenting existing data management and analytics with data streaming and real-time analytics.

**Evaluate the potential of data streaming and real-time analytics.** With a greater selection of open source programs and frameworks as well as the latest generation of commercial tools to choose from, data streaming and real-time analytics are poised to become mainstream, possibly displacing older technologies and practices. TDWI research finds that significant percentages of organizations are using technologies to manage and analyze IoT sensor data, web logs and customer behavior data, mobile and geolocation information, and more. Organizations should develop a strategy for how to augment existing data management and analytics with data streaming and real-time analytics and which business objectives would benefit.

denodo.com

Denodo is a leader in data virtualization providing agile, high-performance data integration, data abstraction, and real-time data services across the broadest 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 realtime business intelligence and self-serviceability.

Founded in 1999, Denodo is privately held, with main offices in Palo Alto (CA), Madrid (Spain), Munich (Germany), and London (UK).

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

research

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

tdwi

**Transforming Data**
**With Intelligence™**

555 S. Renton Village Place, Ste. 700
Renton, WA 98057-3295

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org