playbook

Modernizing Data Integration

By David Stodder

Sponsored by:

StreamSets

tdwi TRANSFORMING DATA WITH INTELLIGENCE™

Introduction

Business conditions and objectives are always changing, which means that data integration cannot stay the same. To gain actionable insights into customer behavior and loyalty, financial performance, process optimization, and opportunities for new products and services, decision makers need timely, trusted, and complete data. However, digitally transformed applications, new online channels, and smart mobile and sensor devices are generating enormous and diverse volumes of data. Provisioning users with the right data has never been more challenging.

TDWI finds that organizations have a strong interest in modernization. In just-published TDWI research, more than half of organizations surveyed are actively modernizing data platforms, integration, and management (54 percent) and 29 percent are planning to modernize in the near future.[1] Modernization enables organizations to grow their use of data, innovate, increase efficiency, and improve user satisfaction. It also involves updating data governance to keep pace with data integration practices and transforming the organization's culture to support data-driven objectives.

Data democratization is a major cultural change that data integration modernization must support. It is about empowering a wider spectrum of people to use data effectively, which includes providing them with self-service data integration capabilities. Many users want to be more independent of IT in selecting data sources, developing data pipelines, and setting up ETL processes to prepare data for business intelligence (BI) reports, dashboards, applications, and analytics. In TDWI research, 30 percent say that adding self-service functionality is extremely important to their data modernization strategy and 43 percent say it is very important.[2]

[1] See Q3 2022 *Best Practices Report: Maximizing Business Value with Data Platforms, Data Integration, and Data Management*, Figure 1, online at tdwi.org/bpreports

[2] Ibid., Figure 4.

Organizations need a strategy for avoiding data chaos. The danger of data democratization is that piecemeal development and deployment of data integration jobs can lead to data processing and networking performance problems as well as data duplication and inconsistency. Users become frustrated by having to spend so much of their time on data integration. They suffer delays in getting the data they need due to having to repair and rerun faulty jobs.

Attaining the right balance between self-service data integration and IT-managed enterprise resources is key to enabling users to be productive in achieving data-driven business goals. With the right balance, organizations can improve scalability, avoid data chaos, and gain faster speed to insight.

Today's data requirements demand different solutions. New use cases and workload types put pressure on organizations to have a broader selection of data integration options. Most processes, such as ETL, use batch processing. However, many users, analytics models (including AI/ML), and data-driven applications demand faster processing that can deliver near-real-time data, if not true real-time data streams. Organizations need to match the right technologies to the appropriate use cases. As they adopt near-real-time or real-time data integration technologies, they need to prioritize which workloads require the most frequently updated data and real-time data streams.

Finally, migrating data management and data integration to the cloud is a common objective. Many organizations execute phased migrations, moving selected data and integration jobs to the cloud over time. Data localization and residency laws may require that organizations collect, process, and store data about a nation's citizens or residents in the respective countries. As a result, hybrid environments are common; these feature a combination of on-premises and cloud-based data. Organizations are also storing and processing data on multiple cloud platforms. The resulting hybrid multicloud data environments add complexity to data integration.

With these trends and challenges in mind, we now consider some of the top business cases driving data integration modernization and how these shape organizations' strategies.

Business Drivers for Modernization

It is easy to get lost in the technical issues of data integration, but the ultimate reason organizations need fast, scalable, agile, and resilient data integration is to serve business objectives. The success of initiatives such as digital transformation of business processes rests on how fully organizations can realize value from new and diverse data sources for analytics, operational efficiency, and business integration. Organizations need to overcome limitations of legacy systems so they can move forward with data-driven business objectives.

A prominent business driver is to improve customer engagement. Data integration processes must supply continuous data for analytics and AI/ML to develop deeper insights into customer preferences and behavior. Rapid expansion in e-commerce, portals, websites, and social media has given firms in a variety of industries (not just retail) the ability to engage directly with customers.

Engagement across channels, including on mobile devices, generates voluminous and varied data that organizations are interested in integrating for real-time monitoring and analytics. Organizations need accurate, complete, and timely data to drive customer personalization. Timely data helps decision makers spot trends that suggest they need to make changes immediately to marketing, engagement, and product development strategies to gain the best outcomes.

Customer data trends and patterns are also helpful to executives in other operations (such as fulfillment and logistics) to improve planning. Corporate executives can use the insights to determine how best to grow the business. Agile data integration systems are able to adjust collection, ingestion, and processing steps to fit different users' needs.

Organizations that may not have enough in-house data for understanding customer and consumer demand, market directions, and the competitive landscape need data integration systems that enable users to ingest data from external sources to complete or enrich internal data. External data could come from data brokers, supply and demand chain partners, or data marketplaces and exchanges.

Integration of external data is easier due to broad adoption of standard application programming interfaces (APIs). Organizations can use APIs to locate and access information on other organization's servers; conversely, you can make your data available to other firms. Integrating external data with internal data can offer advantages, but it also adds to existing complexity challenges of data quality, consistency, governance, and the reliability of sources.

Additional business priorities that TDWI research finds are driving modernization initiatives include:

- **Increasing operational efficiency and effectiveness.** Data integration needs to fuel BI reports, dashboards, and analytics so managers can optimize business processes, reduce costs, and improve performance. Modernization must remove data integration and data pipeline bottlenecks that delay getting the right data to the right people at the right time. Increasingly, line-of-business (LOB) and operational managers can benefit from near-real-time or real-time data feeds in dashboards or for real-time analytics so they can address situations and trends quickly.

- **Generating new business strategies and models using analytics.** Recent years have brought significant and often unexpected changes to marketplaces, customer behavior, supply chains, and the economic landscape. Inflexible data integration makes it difficult for organizations to be agile; they cannot respond to changes and seize new opportunities. Executives and managers need to examine data from different perspectives and use numerous variables to uncover insights.

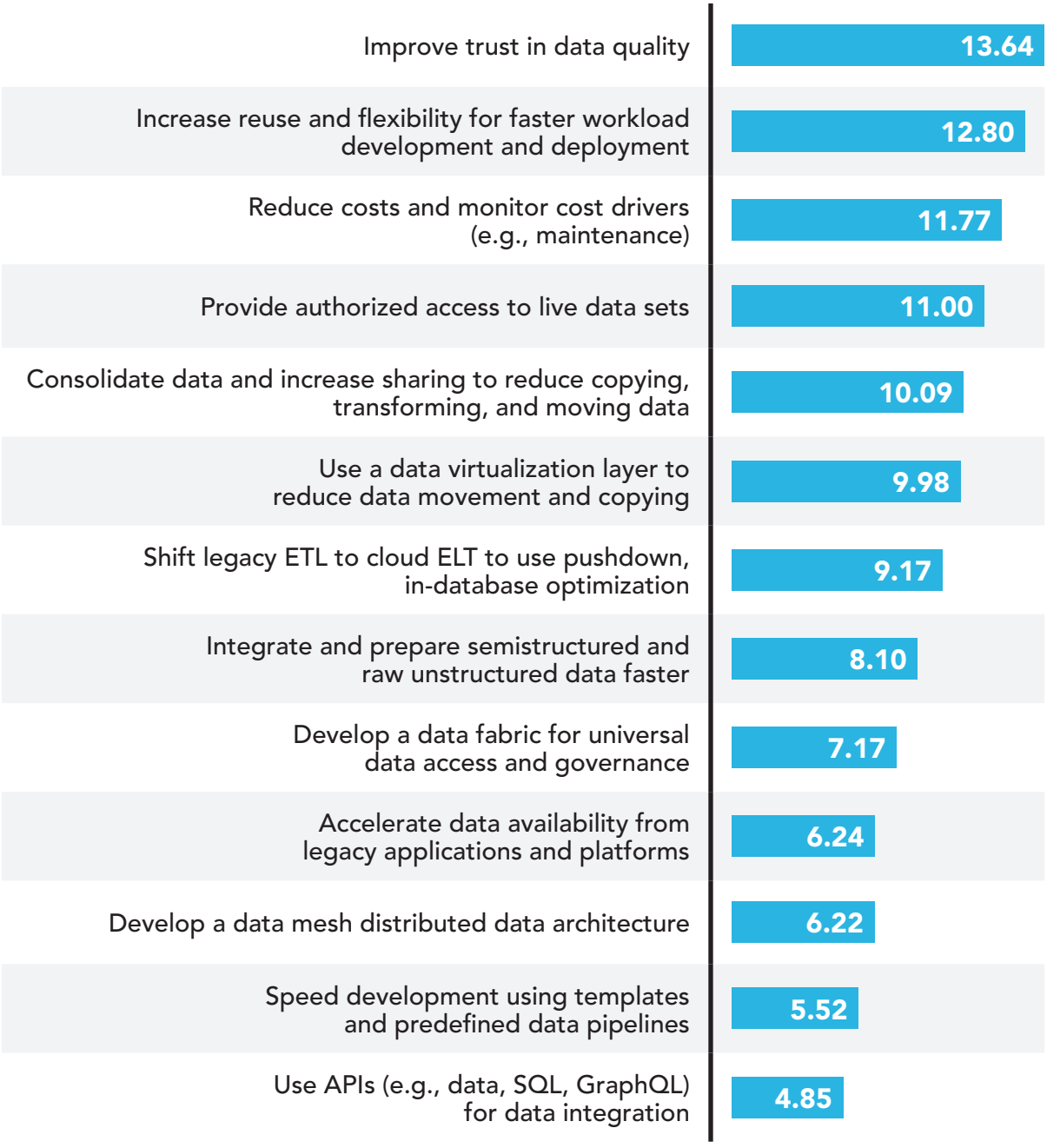
Organizations need scalable data integration so that data scientists and engineers can deploy data pipelines quickly and easily to provision data for analytics and AI/ML.

- **Supporting faster (including automated) decision-making.** In TDWI research, organizations surveyed show strong interest in modernizing data integration to eliminate latency that thwarts timely, data-informed decisions, including automating decisions inside applications. To accomplish this objective, organizations are tightening connections between analytics and applications. AI-infused automation often depends on real-time data streaming to feed algorithms that ensure applications respond to current data about situations, trends, and behavior.

Critical Plays for Getting Started

To modernize data integration, you need a strategy: an overarching plan that unifies decisions about changing technologies, processes, and practices. Figure 1 shows how organizations surveyed by TDWI rank modernization objectives. In this section, we will highlight five recommended plays for getting started with modernization that address many of these objectives.

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



Source: TDWI Q3 *2022 Best Practices Report: Maximizing Business Value with Data Platforms, Data Integration, and Data Management*

Figure 1. *Based on answers from 316 respondents who were asked to rank the options. Ordered by weighted average.*

FOCUS ON IMPROVING DATA TRUST

This is the top-ranked objective in our research. To merit users' trust, data needs to have fit-for-purpose data quality, currency, accuracy, and completeness. Operational dashboards, for example, have stringent data quality and accuracy standards, but data pipelines for data science projects may demand raw, real-time data feeds or that only specific and limited data quality and enrichment measures be taken.

Data governance is also important to building trust. You need to know that your data does not violate rules and policies regarding the collection, analysis, and sharing of sensitive data. Ensuring only authorized access to data ranked fourth in Figure 1. Access control is an element of trust vital to your organization's data security and to guiding users so they do not violate data privacy regulations.

IMPLEMENT DATAOPS FOR SPEED, SCALABILITY, REUSE, AND FLEXIBILITY

In Figure 1, we can see the importance of reuse and flexibility for faster workload development and deployment. This collection of objectives ranked second. Implementing DataOps can help you achieve such goals as you deploy possibly hundreds or thousands of data pipelines and transformation processes.

DataOps borrows from agile and DevOps methodologies to give you a framework for eliminating delays and inefficiencies in often disconnected phases of data life cycles, including development, data collection and integration, data quality, profiling, transformation, enrichment, governance, and capacity planning. DataOps has proven helpful to organizations as they orchestrate data pipelines, improve efficiency to contain costs, and apply modern technologies for end-to-end data integration.

DataOps also promotes fuller use of metadata resources such as data catalogs in the development and operationalization of data pipelines. The data intelligence potentially managed by data catalogs can make

it easier to improve data quality, shorten the task of locating and collecting relevant data, and ensuring data governance.

IMPROVE AND AUTOMATE MONITORING

The objective ranked third in Figure 1 is to reduce costs and monitor cost drivers. To mitigate risks, you need good visibility into data integration processes through monitoring. A good strategy is to set up metrics and dashboards for tracking problems and measuring success in meeting performance and data provisioning goals. Automation—including using the latest visual interfaces and AI-driven capabilities—helps you avoid costly and time-consuming manual coding, testing, and performance monitoring of data pipelines and transformation jobs.

Automated monitoring tools can ensure that data pipelines and ETL processes stream or load all the intended data into target destinations on time, and notify data engineers and managers when there are problems. Automation in solutions helps you spot problems quickly and increase consistency in data integration jobs, which is important to reducing risks in the most complex data sourcing and transformations. Monitoring and testing uncovers, for example, changes in data formats over time, errors in key relationships, and other hazards that often lead to unsuccessful data integration jobs. Monitoring enables you to catch problems before bad data moves downstream into BI reports, dashboards, and analytics.

ADDRESS REQUIREMENTS FOR DATA STREAMING AND REAL-TIME ANALYTICS

For most use cases, reducing data latency adds significant business value. Data streaming has become mainstream, including through use of open source Apache Kafka and related technologies for stream processing. Streaming enables data to arrive continuously from thousands of sources such as IoT sensors, log files, financial trading systems, social networks, mobile devices, and in-game player activity. You may choose to land streaming data in a data lake or perform analytics on streaming data in motion.

TDWI research finds that many organizations struggle with integrated analysis of streaming data and historical data. Many also have difficulty choosing between data streaming, change data capture (CDC), fast-batch processing, and other methods of reducing data latency for their projects. There is no one-size-fits-all answer.

As the spectrum of use cases broadens from standard reporting on one end to real-time analytics on the other, you should expand your palette of data integration solutions. Carefully assess requirements and perform proof-of-concept testing with data samples to determine which is the best fit and delivers the most business value.

UNIFY DATA INTEGRATION OPTIONS TO INCREASE EFFICIENCY AND REDUCE SILOS

Having numerous data silos creates data integration challenges. In Figure 1, organizations rank consolidating data fifth among objectives. Many organizations are choosing to consolidate data silos into a single cloud data warehouse or data lake. However, with hybrid, multicloud data environments common and data democratization tending to spawn new data silos, many organizations need a data strategy that favors consolidation where possible but also addresses distributed data integration.

Unifying data integration technologies and processes is helpful. Data silos often come with siloed data integration processes that are not reusable and may not be consistent across your organization. Integrating technology options more coherently into a single holistic system would help you assign the right option to each use case rather than be limited to what is available for each data silo. For example, it would be easier to determine if standard ETL fits user requirements or if it would be better to go with the variant of loading data into the target data platform first and then transforming it (ELT).

Unified data integration systems can also unify monitoring. This relieves data engineers and other users from having to go from one tool's monitoring interface to another. Organizations are able to gain a holistic view of whether each data integration option is provisioning

data as intended. You would be able to eliminate jobs that are causing problems or are no longer valuable. Unified visibility enables organizations to reduce duplication, increase reuse, and expand deployment of modern technologies where appropriate to increase productivity and accelerate business value.

Conclusion: Modernization Is Essential for Success

You need to move past legacy technologies, processes, and practices to achieve ambitions for more data-driven operations, data-rich applications, and wider development and deployment of analytics. Automated AI-infused data integration tools can improve scalability and reduce errors. With modern tools, users, data engineers, and IT developers can replace the slow, manual processes that have dominated legacy systems and practices.

Modernized, automated technologies and practices are essential as democratized users in LOBs and operations exercise more control over data integration. Self-service users need ease of use, built-in guidance (often through wizards), and predefined routines that reduce routine manual work. Fresher data is usually more valuable, so organizations need to develop strategies that align the latest technologies for reducing data latency, including real-time data streaming, with use cases where speed increases value.

Meet Top Data Integration Modernization Objectives

(Content provided by StreamSets)

StreamSets helps data teams easily start meeting the top data integration modernization objectives uncovered by TDWI. Here's how:

Critical Steps to Get Started	How StreamSets DataOps Platform Makes It Easy
Focus on improving data trust	Data SLAs and rules enable you to expose hidden problems in your data flows, create guardrails and quality checks, and then manage by exception.
Implement DataOps for speed, scalability, reuse, and flexibility	StreamSets DataOps Platform is built for scalability, reusability, and flexibility: • Pipeline fragments make it easy to capture, reuse, and refine business logic, encapsulating expert knowledge in portable, shareable elements • Dynamic pipelines let you ingest more data without building more infrastructure, and different teams can innovate at their pace and without any repercussions to the data engineering team • Python SDK enables templatizing data pipelines for scale so you can create hundreds of pipelines with just a few lines of code. See it in action. • "Mission control" across all your environments enables you to move between clouds and on premises at the press of a button.
Improve and automate monitoring	StreamSets "mission control" panel and topologies show you how systems are connected and data flows across the enterprise. Automatically see when a new integration point is created or a more direct data route is available. In addition, you can know where data comes from to understand and explain outcomes—for example, in AI/ML models.

(Table continues on next page)

Critical Steps to Get Started	How StreamSets DataOps Platform Makes It Easy
Address data streaming and analytics needs	StreamSets was built to stream data for real-time analytics and real-time smart applications (i.e., fraud prevention, cybersecurity, making real-time offers).
Unify data integration options	StreamSets lets you build batch, CDC, ETL, ELT, and ML pipelines—and all from a single UI. The "mission control" panel lets you seamlessly integrate data from any system, on premises, cloud, hybrid, or even mainframe.

Leading Enterprises Trust StreamSets

Some of the largest companies in the world trust StreamSets to power millions of data pipelines for modern analytics, AI/ML, smart applications, and hybrid integration.

See how these and other leading companies have modernized data integration with StreamSets

The StreamSets data integration platform provides:

Request a demo to see StreamSets DataOps Platform in action

About Our Sponsor

StreamSets

At StreamSets, a Software AG company, our mission is to ensure that organizations and data teams thrive in today's world of constant change. We do this by embedding the DataOps philosophy of "continuous data for the connected enterprise" into the StreamSets DataOps Platform. StreamSets empowers your data team to build, run, monitor, and manage smart data pipelines for the modern data ecosystem.

StreamSets provides a single design experience for all design patterns for 10x greater developer productivity; smart data pipelines that are resilient to change for 80 percent less breakages; and a single pane of glass for observing and monitoring all pipelines to eliminate blind spots and control gaps. With StreamSets, you can deliver continuous data for the modern data ecosystem and hybrid integration in a world of constant change.

About the Author

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

About TDWI Research

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

About TDWI Playbooks

TDWI Playbooks provide data professionals with a summary of important key factors about contemporary data-related topics. Playbooks present the issues and challenges facing enterprises about each topic and offer a concise list of proven best practices to succeed in a particular area of analytics, business intelligence, or data management. Playbooks are written by TDWI research analysts and faculty who synthesize their research and experience into easy-to-understand explanations and practical recommendations that enable data professionals to apply the best, most productive approaches and techniques to their projects or initiatives.

tdwi
Transforming Data With Intelligence™

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

E info@tdwi.org

tdwi.org