TDWI INSIGHT ACCELERATOR

Turning Traditional ETL Systems into Intelligent DataOps Pipelines

By James Kobielus

Sponsored by

aws | wavicle DATA SOLUTIONS

tdwi

OPERATIONAL PAIN POINTS WITH TRADITIONAL ETL

ETL systems are the backbone of data analytics in any modern organization. They *extract* data from transactional applications, *transform* it so that it's suitable for analysis in downstream applications, and *load* it into data warehouses, data marts, and other key data repositories.

Traditionally, the ETL process has involved building structured workflows—also known as pipelines—that automate batch processing and movement of data between sources and downstream repositories, applications, and users. The key pain points with traditional ETL and other data-integration systems include:

- **COST.** Traditional ETL and data integration systems are expensive to acquire, operate, and maintain, having been manually built and often managed through labor-intensive processes.
- **COMPLEXITY.** Traditional ETL and data integration systems do not easily connect to existing infrastructure components. Considering the growing complexity of modern data infrastructure, this often results in IT professionals having to develop custom ETL code. Making this chore even more difficult, traditional ETL systems provide a limited set of transformations, making it difficult to develop bespoke data integration flows.
- **CONSTRAINTS.** Traditional ETL systems are engineered primarily to process structured, tabular data and have limited capabilities for handling the semistructured and unstructured data generated and consumed by more applications. Traditional ETL systems cannot support low-latency workloads and are unsuited to the growing range of real-time and streaming applications upon which modern business

***Figure 1**. Based on answers from 204 respondents.*

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

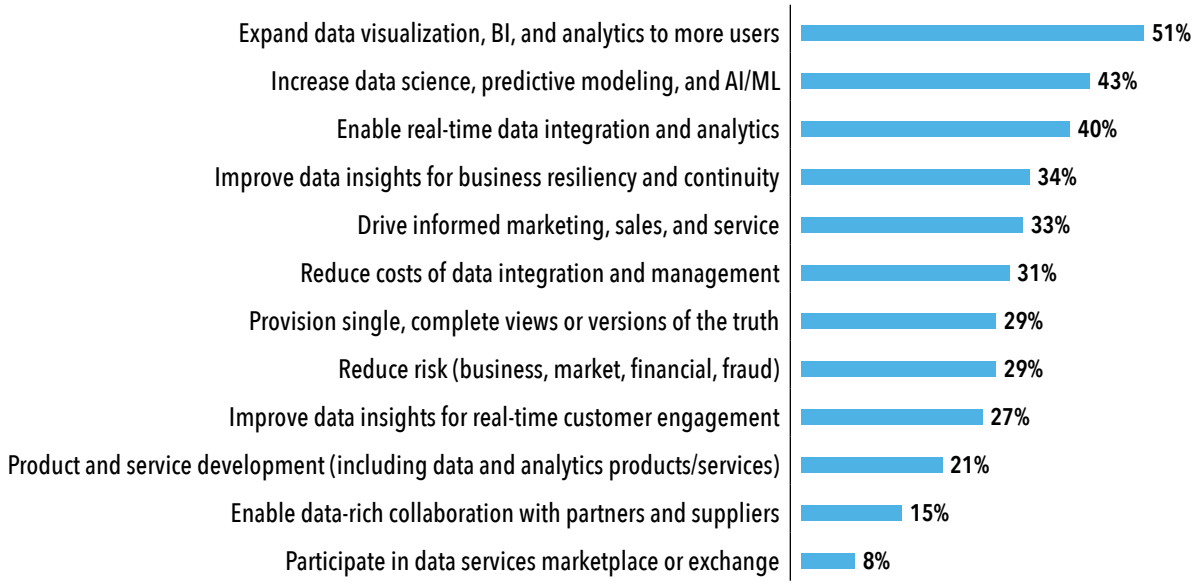

Figure 2: *Based on answers from 437 respondents.*

depends. They execute on fixed processing nodes and are not engineered to scale with sudden workload spikes. Stringent ETL software licensing terms make it difficult to customize the platform for business requirements. Also, traditional ETL systems cannot easily track changes made by users, making it difficult to find the root cause of errors.

As attested to by data management professionals in a recent TDWI survey and shown in Figure 1, another key pain point of traditional ETL is a shortage of trained, certified personnel.[1] This, coupled with demands on them to perform largely manual, time-consuming functions along with a persistent lack of collaboration with other data integration stakeholders, can create significant bottlenecks and delays for a DataOps staff. That staff will be hindered in its ability to revise ETL code to incorporate changes in data sources or accommodate new downstream demands for specific data elements in reports, dashboards, and other analytics.

BUSINESS IMPERATIVES DRIVING ETL MODERNIZATION

ETL ultimately serves the business. For modern organizations, success rests on how well an organization manages ETL processes within the larger DataOps and MLOps pipelines under which it operationalizes data and analytics.

Any modern ETL system should be able to cleanse, augment, and enrich all types of data for a growing range of sophisticated applications. As reported by data management professionals in a recent TDWI survey and shown in Figure 2, the

[1] Source: 2022 *TDWI Best Practices Report: Unifying Data Management and Analytics Pipelines*, online at tdwi.org/bpreports.

key objective when modernizing ETL and other data integration systems is to improve how well various data-intensive applications—especially business intelligence, predictive modeling, machine learning, artificial intelligence, and real-time analytics—serve the bottom line.[2]

When planning to modernize their ETL systems, organizations should keep the following business requirements uppermost:

- **COMPLIANCE.** Organizations require ETL and other DataOps solutions that support enterprise-wide data governance policies and help them comply with data privacy and industry-specific data use regulations.
- **AGILITY.** Business is a combination of the routine and the unexpected. To address the full spectrum of requirements, ETL and other DataOps functions must support both repeatable orchestration and ad hoc workloads, enabling enterprises to respond to immediate needs.
- **SIMPLICITY.** Businesses perform best when the full spectrum of users is empowered with data analytics. This requires that ETL and other DataOps systems be visual, self-service, free of coding, and easy to use. It should simplify consumption of business data, sparing users from the confusion created by low-quality, ill-defined, and inconsistent data.
- **AUTOMATION.** Data integration is a thankless chore that, fortunately, can be automated. Modern ETL solutions should relieve users from the need to engage in unnecessary, manual, time-intensive data integration chores. To reduce the time and cost necessary to build, deploy, and manage workflows, ETL solutions should be tailored to the needs of particular DataOps stakeholders, such as data engineers and data scientists.

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

THE JOURNEY TO ETL MODERNIZATION

Undertaking ETL modernization can be a daunting process. It can involve a wide range of overlapping and interdependent technical, operational, and business issues.

As enterprise DataOps professionals plan their ETL road maps, they should consider a target architecture that incorporates the following pillars of a modern DataOps pipeline:

- **UNIFIED.** ETL pipelines should run workloads in a unified DataOps architecture that spans hybrid cloud, multicloud, and other complex topologies. Furthermore, they should integrate smoothly with machine learning operations (MLOps) pipelines for more efficient data preparation and training of ML models into intelligent applications.
- **FLEXIBLE.** ETL pipelines should support flexible scaling of DataOps workloads. They should elastically provision computational, storage, and other cloud resources as needed. They should be able to ingest and process data from structured, semistructured, and unstructured sources. Pipelines should enable enterprises to build and execute any type of ETL flow, including those involving data cleansing, profiling, monitoring, cataloging, and preparation. Finally, they should enable creation, orchestration, and management of multiple ETL pipelines in parallel, as well as reordering execution of ETL workloads as needed.
- **GOVERNED.** ETL pipelines should support the full range of governance workloads for all data assets ingested, prepared, and delivered to downstream applications. They should integrate with the data catalog, data profiling, data cleansing, data lineage, master data management, metadata management, and business glossary infrastructure for all domains.

- **ACCELERATED.** ETL pipelines should be optimized for real-time, low-latency, continuous processing. They should be able to run migrated workloads in a distributed, in-memory, cloud-native architecture that natively supports Spark, Flink, Kafka, and other stream computing backbones.
- **OBSERVABLE.** ETL pipelines should be monitored from end to end with intelligent data observability tools. This is important for anomaly detection, predictive issue detection, and closed-loop remediation of problems with ETL pipelines and their handling of specific workloads.
- **INTELLIGENT.** Within a modern DataOps architecture, an ETL pipeline should adapt dynamically to changing contexts, workloads, and requirements. This involves the embedding of machine learning (ML) intelligence at every pipeline node and in every process. The architecture should be able to adapt to cross-pipeline dependencies based on data, time, and conditions. It should be able to automatically discover new and revised data assets to be extracted into the pipeline; validate data as it is ingested into the pipeline; adapt its logic to new sources, contexts, and processing requirements; and pre-emptively remediate technical, workload, and performance issues before they become showstoppers. Finally, an intelligent ETL pipeline should automatically generate real-time contextual recommendations that guide DataOps professionals in managing, optimizing, and troubleshooting workflows and jobs of various degrees of complexity.

RECOMMENDATIONS

Evolving a traditional ETL platform into an intelligent DataOps pipeline is a key component of cloud modernization.

Migrating to a modern ETL platform can be a complex undertaking. It requires a smooth transition of existing ETL workflows to the new platform without disrupting business or technical operations. Key steps in this journey include:

- **PLANNING.** Enterprise DataOps professionals should use the criteria discussed in this report when choosing the best ETL modernization target architecture for their needs. They should consider whether available solutions are well suited to the target environment—on-premises, public cloud, or some hybrid or multicloud configuration—in which the target ETL pipeline will run. In planning a migration to the target architecture, they should also identify the best solutions for automating as much of this work as possible. Once the chosen solution has been identified and migration tools selected, ETL professionals should prioritize which ETL jobs will be moved (and in what sequence) to the target environment.
- **IMPLEMENTATION.** ETL migration usually involves moving multiple workflows and associated jobs. DataOps professionals may choose to rebuild ETL workflows from scratch on the target platform. They may choose to replicate exact copies of workflows from the legacy environment onto the target environment. Alternatively, they may automate the conversion of existing ETL workloads from traditional tools to the target platform. DataOps professionals should ensure that the movement of ETL workloads to the target environment does

not impact data availability to downstream applications, users, and stakeholders. After the new ETL environment has been activated, it should be validated thoroughly to ensure it works reliably with all existing ETL jobs.

- **OPTIMIZATION.** ETL migration to the target environment need not be purely lift-and-shift. New ETL workflows can and should be developed to take advantage of the scalability, performance, and other advantages of the new platform. If nothing else, redundant ETL flows should be combined to save processing, storage, and bandwidth resources.

The journey to ETL modernization requires that enterprises migrate to an elastic, fully managed, cloud-native DataOps infrastructure. Ideally, this infrastructure should be unified with cloud-based MLOps pipelines and data lakehouses to support building, training, and deployment of the data-driven intelligent apps upon which modern business depends.

ABOUT OUR SPONSORS

Amazon Web Services provides a comprehensive and broadly adopted cloud offering. Over the past 15 years, AWS has been continually expanding its services to support virtually any cloud workload, and it now has more than 200 fully featured services for compute, storage, databases, networking, analytics, machine learning and artificial intelligence, Internet of Things, mobile, security, hybrid, virtual and augmented reality, media, and application development, deployment, and management from 81 Availability Zones within 25 geographic regions, with announced plans for 21 more Availability Zones and seven more AWS Regions in Australia, India, Indonesia, Israel, Spain, Switzerland, and the United Arab Emirates. Millions of customers—including the fastest-growing start-ups, largest enterprises, and leading government agencies—trust AWS to power their infrastructure, become more agile, and lower costs. To learn more about AWS, visit aws.amazon.com.

Wavicle Data Solutions provides award-winning data and analytics consulting services that help businesses across multiple industries leverage modern data architecture and cloud-based technology to capture, analyze, and act on their growing volumes of data. Our solutions reduce the time, cost, and risk of clients' projects while improving the quality of their data and insights. Trusted for our technical expertise, business-led approach, and partnership focus, our clients rely on us to solve their most complex business issues quickly and cost-effectively. Our solutions include data management, cloud engineering, data visualization, and data science and analytics. Wavicle has been recognized by *Inc.* 5000 as one of the fastest-growing private companies in America in 2021, 2020, and 2019 and by *Crain's Chicago Business as* one of Chicago's fastest-growing companies in 2021 and 2020. We were also named a *Chicago Tribune* Top Workplace in 2020 and 2021.

Learn more at wavicledata.com.

ABOUT THE AUTHOR

James Kobielus is senior director of research for data management at TDWI. He is a veteran industry analyst, consultant, author, speaker, and blogger in analytics and data management. He focuses on advanced analytics, artificial intelligence, and cloud computing. Kobielus has held positions at Futurum Research, SiliconANGLE Wikibon, Forrester Research, Current Analysis, and the Burton Group and also served as senior program director, product marketing for big data analytics, for IBM, where he was both a subject matter expert and a strategist on thought leadership and content marketing programs targeted at the data science community. You can reach him by email (jkobielus@tdwi.org) on Twitter (@jameskobielus) and on LinkedIn (https://www.linkedin.com/in/jameskobielus/).

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.