playbook

Future-Proofing Your Data Pipeline

By David Loshin

Sponsored by:

StreamSets
A SOFTWARE AG COMPANY

tdwi TRANSFORMING DATA WITH INTELLIGENCE™

An Introduction to Data Pipelines

In the early days of computing, organizations were keenly aware of their need for speedy transaction processing, so they prioritized processing over the subsequent use of those transactions. Data warehouses evolved as a means of data segregation; pulling data out of transaction systems and putting it into a separate analytics environment precluded analyses from interfering with transaction performance. Decades of data segregation have imprinted ETL (extraction, transformation, and loading) as a necessary practice in which information technologists and database developers direct what and how data sets are made available for analysis.

In today's data-driven world, there is little patience for barriers to data access, which has inspired many organizations to rethink how new technologies such as cloud computing, open source tools, and sophisticated AI/ML analytics can add value. As enterprise technologists continue to migrate their production information environments to modernized (mostly cloud-based) data management platforms, three trends are influencing sophisticated changes in the management and implementation of the data analytics life cycle:

- The growing interest in allowing a range of data consumers with different levels of technical skills to take advantage of machine learning, AI, and other advanced technologies for data analytics.
- The expansion of the enterprise information environment beyond the boundaries of the traditional data center, incorporating both internally accessible and externally produced data sources to meet organizational analytics demands.
- The radical changes in the ways organizations continued operations during the pandemic, which surfaced new opportunities to improve the customer experience with broader and more sophisticated uses of both internal and externally acquired data resources.

Meeting data consumer expectations while your enterprise is modernizing its data management platforms, adopting a cloud platform, and moving its application systems to that new cloud platform is critical. Yet many organizational data integration architectures remain technically limited to the outmoded lock-step batch data integration model common in most data warehouse architectures. If not carefully considered when migrating technical capabilities to the cloud, traditional data integration patterns can result in an unmanageable, complex environment for data provisioning. Organizations acknowledging the need to streamline the acquisition, management, accessibility, and delivery of a wide array of data resources in an efficient and trustworthy manner are recognizing the value of data pipelines.

In its recent Best Practices Report about data management and analytics pipelines, TDWI defines a data pipeline as "a chain of connected processes that take data, models, program code, and other digital assets and prepares them for delivery to and consumption by downstream applications." As organizations collect ever-increasing volumes and types of data and move to support a growing number of real-time data needs, they must consider how orchestrated data pipelines offer an alternative to traditional data integration. This playbook discusses what a data pipeline is and why it is important. It reveals current and emerging trends in pipelines (including pipelines to support data and analytics, automation in data pipelines, and cloud-native pipelines) as well as steps for getting started future-proofing your data pipelines.

Core Principles: Data Pipelines Support DataOps

Data pipelines provide a fundamental means for deploying and orchestrating a variety of functional components for handling and managing data across the analytics life cycle. Data pipelines can be used to coordinate data engineering functions (such as data asset discovery, acquisition, ingestion, transmittal, transformation, virtualization, and delivery). These functions constitute a DataOps approach employing open source tools, cloud-native capabilities, and data utility frameworks to design, implement, maintain, and orchestrate functionality within a production cloud-based/distributed data architecture. Figure 1, from the same TDWI Best Practices Report, shows the functionality that data pipelines can encapsulate.

Architecture of DataOps and MLOps pipelines in CI/CD workflows.

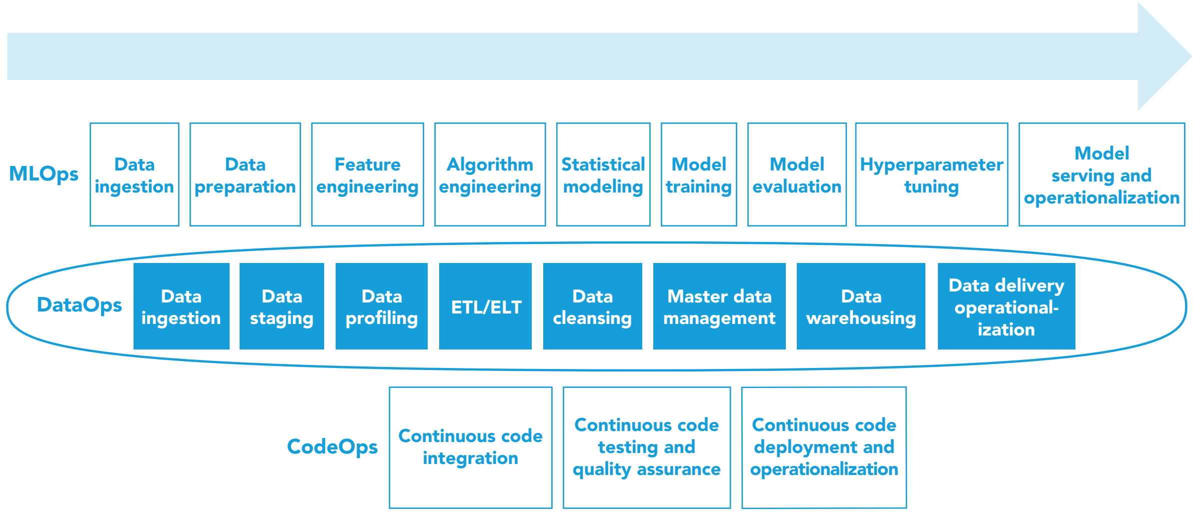


Figure 1. *Data pipelines encapsulate the DataOps components.*

Data pipelines enable DataOps through:

- **Modularization.** Data pipelines implement processing componentry that can be modularized as services; these services enable the creation of distinct workloads that facilitate the movement of data from sources to a dedicated destination.

- **Simplified configuration.** Data pipeline tool vendors typically provide visual tools that simplify the design and creation of pipelines meeting specific data-consumer needs. Bypassing the IT bottleneck empowers data analysts and data scientists to rapidly (and independently) configure, deploy, test, and fix their data engineering processes.

- **Managed orchestration and automation.** A complex organization might have hundreds or thousands of simultaneously executing data pipelines, and manually managing and overseeing these can be an operational nightmare. Data pipeline orchestration platforms can configure many pipelines to run automatically when triggered by events (such as the delivery of an updated data set or the arrival of streaming data).

- **Integrated governance.** Enterprises can devise data pipelines that can employ encapsulated data services supporting the integration of performance, scalability, data quality, and data protection obligations associated with the contexts of usage scenarios and data-consumer privileges and expectations.

Data Pipelines: Current Trends

According to TDWI's research, the main data management and analytics pipeline use cases remain well-acknowledged core applications—despite the promise of advanced applications that can leverage data pipelines. When respondents to the Best Practices Report survey were asked about how they used DataOps pipelines in their organizations, 57% said "data transformation," 52% said "data extraction," and 50% said "ETL/ELT."

Although fewer than 20% of the respondents indicated use cases for data logging, data versioning, data virtualization, and data orchestration, some current trends suggest that the use of data pipelines for a wider variety of DataOps objectives has growth potential. Some organizations are already taking advantage of standardized approaches to data pipeline development and management and data pipeline automation in predictable ways. Current trends include:

Performance computing. Cloud-based data pipelines have several benefits. Aside from providing access to data resources that have been migrated to the cloud, these pipelines are engineered for elasticity and scalability. Computational resources can automatically grow to satisfy the service-level requirements for ingesting and processing many incoming data sources and delivering data to a wide number of data consumers, no matter how great the data volumes grow.

Decentralized data. Data architecture modernization in the cloud allows for separating data management from the computational resources that operate on that data. This eliminates the need for data centralization in a traditional data warehouse, especially as data pipelines facilitate access to any shared data assets in their native forms.

Stream processing and integrated analytics. Organizations can expand beyond traditional static data sources by adapting data pipelines to ingest continuously flowing information streams in real time. Real-time event processing is powered by continuously blending streaming data with integrated analytics models. Examples intended to improve customer experience include real-time fraud detection in streaming financial transactions, improved precision and accuracy in predicting package delivery times, or up-to-date visibility into a customer's account portfolio.

Self-service enablement. Another artifact of the traditional data integration processes that data pipelines can help eliminate is the dependence on IT to configure and develop ETL integration processes. The simplified development environments that data pipeline tools provide encourage data consumers to design, test, and deploy their own pipelines, thereby avoiding the IT bottleneck.

Data Pipelines: Emerging Trends

According to research for the Best Practices Report, 52% of respondents believe that DataOps pipelines need more structuring for data integration, governance, quality, and compliance. Over time, organizations will become more accustomed to deploying and managing their data pipelines using a modern, cloud-based data architecture. Correspondingly, data analysts will better understand the potential of enabling real-time access to a combination of data at rest and data streams. Several emerging data pipeline trends empower data analysts to access the right data assets for their specific business use cases, even with increasing data volume and variety. These trends include:

Modular data transformation building blocks. Enterprises can benefit from the flexibility made possible by self-service configuration. However, too much flexibility will lead to pipeline sprawl. Pipeline sprawl occurs when multiple pipelines funnel data from the same data sources, subject the data to the same (or similar) validations, and perform some of the same transformations, often without the knowledge of the data engineers. An emerging data pipeline trend is to develop modular components embodying declarative transformations that assert a structure or schema for the data emerging from that component. Registering and managing a collection of modules provides two key benefits. First, it provides a palette for data consumers to develop their pipelines. Second, it allows the underlying orchestration manager to identify opportunities for streamlining those data pipelines that employ the same components and reuse the same data streams.

Data product "factory." Archaic methods of ETL data integration cannot satisfy the needs of modernized data use cases. For example, applications supporting a unified customer experience rely on real-time access to trustworthy and current source data. One particular data pipeline trend—the rise of self-service data pipelines—is motivated by a fundamental paradigm shift for enterprise data use. Instead of relying on a preconfigured centralized data warehouse model, data consumers want facilitated, on-demand access to a semantic view of information assembled in real time from decentralized data. Self-service data pipelines free the data consumers to treat the information environment as a data "factory" that can manufacture data products to meet their specific needs.

Integrated governance. Data governance has become more critical as ensuring trustworthy data and protecting sensitive data against unauthorized access and use become equally important. A third concern is fault tolerance—awareness of potential flaws in the production processes for the data life cycle. A continuing trend for data pipeline management and orchestration tools is integrated governance, with the ability to embed data validation and other operational data controls directly into the data pipelines. Outputs of these controls can populate auditing and monitoring services that can identify potential issues and alert data stewards so the issues can be reviewed and addressed.

Critical Plays for Future-Proofing Your Data Pipelines

All organizations that perform analytics already have data pipelines. Our objective, however, is to ensure that your technology plan adequately acknowledges and addresses the emerging trends.

At the highest level, the future opportunities for building and managing a data pipeline are satisfied in a straightforward way by vendors providing drag-and-drop graphical editors for designing pipelines that feed automation and orchestration platforms. Yet when organizations have neither properly prepared their environments nor provided the proper socialization and training, they run the risk of their data consumers vacantly staring at the blank canvas of their data pipeline GUI tools, trying to figure out what they want to accomplish and how they want it to be done.

An appropriately prepared organization will consider the following when future-proofing their data pipelines:

DATA AWARENESS

Empowering data consumers to create the future data product factory requires them to be able to configure their own data selection, engineering, and production pipelines. Future-proofing your data pipelines by raising awareness of the data resources that are at the data consumers’ disposal will increase their likelihood of success. In addition to understanding which data resources are available within an organization, this also requires a detailed description of each data asset's structural, business, and semantic metadata and any restrictions and controls governing which data elements can be accessed (or not). *Data catalog* tools inventory enterprise data assets, and data consumers can search for available data assets best suited to their needs. Access to metadata about data elements and their corresponding data quality constraints helps data analysts identify the data sources to feed a customized data pipeline for producing a curated data stream that meets their specific needs.

DATA CONNECTORS

Future-proof your environment by embracing the modular development processes enabled by low-code/no-code development platforms. Modularity must encompass the two principal aspects of data connectivity: *connector methods* such as JDBC, ODBC, and REST data service connectivity and *dedicated connectors to incoming SaaS data sources* such as Salesforce, Marketo, Eloqua, ServiceNow, and other SaaS data feeds. Data pipeline designers benefit from having a wide array of data connectors.

DATA ENGINEERING AND PRODUCTION

The first aspect of data engineering is often reverse engineering, which uses data lineage tools to survey the existing data landscape and provide a map of the ways data sets are sourced as part of the production process. Reviewing that data informs the data pipeline designer and highlights current data integration processes. In turn, common data standardizations and transformations can be modularized as functional operational components available to the data pipeline designer. Examples include automated data profiling, standardizing case, applying common data corrections, address standardizations, identity resolution, and matching/merging. Data integration functionality powers data production and supports data delivery.

AUTOMATION AND ORCHESTRATION

Future-proof your data pipelines by incorporating automation technologies into your enterprise. Automation means transforming the data pipeline design into code that can be executed automatically within a managed environment. Orchestration is the technology that performs this transformation, automates the scheduling and execution of data pipelines, coordinates dependencies across data pipelines, and optimizes pipelines that share common data sources and integrations.

OBSERVABILITY AND AUDITABILITY

Get in front of the growing wave of data privacy and data protection laws and regulations by future-proofing your data pipelines using integrated governance. This measures compliance with data quality rules, data protection rules, or other types of performance or usability metrics. The data pipeline orchestration should accommodate the collection of interim measures produced at different stages of the data pipelines to provide observability and to allow for visibility into a horizontal audit of pipeline performance.

IMPLEMENTATION

An organization that has considered these facets of data production and utility will be prepared for the future by introducing practical data pipeline development technologies. Once you have established the foundation, an iterative and repeatable process for developing and instantiating data pipelines embraces these steps:

- Determine the appropriate DataOps infrastructure to supplement data pipeline tools.
- Engage the data-consumer communities, solicit their data requirements, identify key usage scenarios, and prioritize data pipelines to develop.
- Reverse engineer existing data flows as a prelude to data pipeline refactoring (potentially identifying new data sources or transformations to introduce).

- Use the data pipeline tool's graphical interface to develop and test new data pipelines. When satisfied with the data pipeline, submit it to the orchestration component for automation and management.
- Maintain an inventory of controls and measures for monitoring the health of the organization's data pipelines and embed those controls for continuous monitoring.
- Maintain an inventory of developed pipelines for data consumers to review and possibly reuse.

Conclusion

Organizations are still acclimating to the concept of DataOps and are still trying to come to terms with the fundamental paradigm shifts enabled by cloud-based data life cycle management and data production. Transforming traditional data integration processes into a data pipeline form is a necessary first step for understanding the benefits of data pipeline orchestration. Plan for the future by adopting technologies that eliminate technical debt and streamline an evolving cloud-based framework for orchestrating a growing array of concurrent high-performance data pipelines.

As your organization becomes more mature in its use of data pipelines for DataOps, it will begin to adapt the pipeline concept to more sophisticated data analytics life cycles that can take advantage of integrated machine learning and other advanced analytics. A holistic approach starts with corporate data literacy and data awareness to level-set the teams and effectively communicate the value of data pipelines. Data pipeline orchestration and management enable your organization to break free of traditional centralized data management paradigms, deploy logical data warehouses and data lakes, and empower data consumers to craft their own data product "factories" that meet their analytics needs.

Ahead of the Curve: Get Started Quickly with the Original Pipelines Built for DataOps

(Content supplied by StreamSets)

Modularization, drag-and-drop GUIs, managed orchestration, and integrated governance—these are the ways TDWI says data pipelines can enable DataOps. However, there's a driving force behind these enablers: the *why of DataOps* that explains the need for modern data pipelines.

The demand for real-time analytics from stakeholders across the enterprise has never been higher. With volatile markets, constant competitive pressure, and continually increasing customer expectations, every organization needs all its employees to make data-driven decisions. Adopting a DataOps approach moves enterprises toward becoming truly connected and data-driven.

Delivering on the promise of continuous data delivery in the face of constant change—a core tenet of DataOps—requires resilient and repeatable pipelines, and that's where StreamSets comes in.

StreamSets addresses the trends and critical plays discussed in this playbook. Here are some examples:

Trends and Recommendations	How StreamSets Stacks Up
Stream processing for real-time analytics with cloud-based data pipelines	StreamSets was built to stream data for real-time analytics and real-time smart applications (e.g., fraud prevention, cybersecurity, making real-time offers). It's a cloud-based platform designed for scalability and elasticity that serves enterprise clients such as IBM, Humana, Shell, and Deluxe.

Trends and Recommendations	How StreamSets Stacks Up
Data decentralization	StreamSets supports data decentralization by allowing you to decouple data storage from integration, making destination changes a simple push of a button. Use Azure one day and AWS the next, or quickly switch between on-premises and cloud warehouses such as Snowflake.
Self-service enablement	Dynamic pipelines let you ingest more data without building more infrastructure. StreamSets' low-code/no-code interface lets the people who know the data best build their own pipelines. Teams can innovate at their own pace without additional development time from the data engineering team.
Modular data transformation building blocks (and more self-service enablement)	Reusable pipeline fragments empower your whole data team to use functionality your data engineers design. Let anyone easily add the same processing logic to multiple pipelines ensuring that they use the logic as designed, maintaining uptime and simplifying pipeline management. Job templates hide the complexity of job details from business analysts. Data engineers define the job details so analysts can start job instances from the templates by modifying pipeline parameter values only.
Integrated governance, observability, and continuous monitoring	StreamSets' mission control panel and topologies show how systems are connected and how data flows across the enterprise. Data SLAs and rules enable you to expose hidden problems in your data flows, create guardrails and quality checks, and manage by exception.
Automation and orchestration	Python SDK enables templatizing data pipelines for scale so you can easily create hundreds of pipelines with just a few lines of code. StreamSets' SaaS-based Control Hub lets you automate and orchestrate pipelines all in one place—schedule your pipelines and receive alerts about data drift or other problems.

Trends and Recommendations	How StreamSets Stacks Up
Data connectors	StreamSets provides flexibility for data engineers to get data from different sources, formats, and databases and lets them easily move it to a wide variety of destinations. With 50 predefined processors, you can meet 99% of your analytics requirements out of the box. Plus, give your pro-level users the ability to include custom code and deliver it as a new element that can be easily reused.

The largest companies in the world trust StreamSets to power millions of data pipelines for modern analytics, data science, smart applications, and hybrid integration.

IBM Humana deluxe trusted business technology gsk RingCentral

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

StreamSets' data integration platform provides these valuable benefits:

Eliminate Data Integration Friction
10x faster delivery

Insulate Pipelines From Unexpected Shifts
Avoid 80% of breakages & rework

Enable Innovation with Centralized Guardrails
Visibility and control across new and legacy environments

Any Platform, Anywhere
Run any workload anywhere

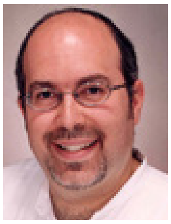
[Request a demo to see StreamSets in action.]

About Our Sponsor

StreamSets
A SOFTWARE AG COMPANY

StreamSets, a Software AG company, eliminates data integration friction in complex hybrid and multicloud environments to keep pace with need-it-now business data demands. Our platform lets data teams unlock data—without ceding control—to enable a data-driven enterprise. Resilient and repeatable pipelines deliver analytics-ready data that improves real-time decision-making and reduces the costs and risks associated with data flow across an organization. That's why the largest companies in the world trust StreamSets to power millions of data pipelines for modern analytics, data science, smart applications, and hybrid integration.

About the Author



David Loshin, president of Knowledge Integrity, Inc., (www.knowledge-integrity.com), is a recognized thought leader and expert consultant in the areas of data management and business intelligence. David is a prolific author regarding business intelligence best practices as the author of numerous books and papers on data management, including *Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph* and *The Practitioner's Guide to Data Quality Improvement*, with additional content provided at www.dataqualitybook.com. David is a frequently invited speaker at conferences, web seminars, and sponsored websites and channels. David is also the program director for the Master of Information Management program at the University of Maryland's College of Information Studies.

David can be reached at loshin@knowledge-integrity.com.

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

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.