CHECKLIST 2022

Modernizing Your DataOps Pipeline to Address Fresh Challenges: Six Best Practices

By James Kobielus

gathr.ai

tdwi TRANSFORMING DATA WITH INTELLIGENCE™

Modernizing Your DataOps Pipeline to Address Fresh Challenges: Six Best Practices

By James Kobielus

Success in the modern economy depends on an enterprise's ability to deliver high-quality data analytics into production applications.

Structured data engineering processes ensure that data and analytics are always accurate, relevant, and fit for purpose. Modern data engineering processes—also known as DataOps pipelines—continuously integrate, transform, and prepare data for production deployment.

In order to keep pace with fast-changing business requirements, enterprises need to address such key technical challenges as:

- Migrating DataOps pipelines to modern cloud infrastructures
- Enabling centralized visibility over the end-to-end DataOps pipeline
- Optimizing DataOps processes for both low-latency and batch processing

Best practices for modernizing your DataOps pipelines:

1. Define a strong business justification for DataOps modernization
2. Identify priority use cases for a modernized DataOps pipeline
3. Align DataOps modernization with strategic cloud data platform implementation
4. Make the necessary investments in enabling infrastructure, tools, and skills for DataOps modernization
5. Bring simplicity into the modernization of the DataOps pipeline
6. Restructure the DataOps pipeline in the process of modernizing it

- Scaling DataOps pipelines elastically to support change data capture; data ingestion; extract, transform, and load (ETL); and other mission-critical workloads

- Orchestrating multiple interdependent DataOps workflows while ensuring that data has been processed and moved as required from sources to target systems

This TDWI Checklist provides DataOps practitioners and other enterprise stakeholders with six best practices for addressing these challenges successfully within a cloud-focused modernization program.

1 Define a strong business justification for DataOps modernization

Success in today's business world depends on how well an enterprise operationalizes data analytics applications.

A modern DataOps pipeline transforms, cleanses, augments, and enriches all types of data. When justifying the modernization of their DataOps pipelines, organizations should lead with the following key points:

- **Consumability:** Businesses perform best when the full spectrum of users is empowered with data analytics. A modern DataOps pipeline can simplify consumption of business data, sparing users from the confusion created by low-quality, ill-defined, and inconsistent data. Key enablers for this include visual, self-service, and no-code tools that streamline complex processes and thereby augment data engineers' productivity.

- **Differentiation:** Business success is all about competitive differentiation. DataOps exists to ensure the continued quality, relevance, and actionability of enterprise data assets. Modernizing the DataOps pipeline can stimulate development of innovative data applications that incorporate conversational AI, deep learning, and streaming analytics. These applications—applied both to back-office processes and customer-facing touchpoints—can help enterprises gain a competitive advantage in today's dynamic economy.

- **Scale:** Business is sustainable when an enterprise's DataOps pipeline attains economies of scale. A modern DataOps pipeline unites siloed data integration, engineering, and governance processes into a scalable, efficient high-performance cloud environment. It provides the bandwidth, memory, processing, storage, and other resources needed to handle expected workloads while meeting latency and service-level requirements. It also makes sure that cloud capacity is used to maximum efficiency.

- **Agility:** Business is a blend of standard operating procedures and fast responses to changing circumstances. To address the full spectrum of requirements, a modern DataOps pipeline must support repeatable, intricate, and interdependent workload orchestrations as well as ad hoc data integration and governance jobs that respond to unanticipated demands. In addition to structured relational data, the modern DataOps pipeline can ingest, transform, cleanse, and deliver diverse data types from new and existing sources,

especially unstructured and semistructured data from the Internet of Things, social media, customer channels, big data, and web apps. It can also support the data ingestion and preparation required for the building, training, and operationalization of machine learning (ML) and other advanced analytics applications.

- **Automation:** Modern DataOps pipelines relieve users from the need to engage in unnecessary, manual, and time-consuming data integration chores. To reduce the time and cost necessary to build, deploy, and manage workflows, modern DataOps pipelines can be designed to automate many of the repeatable functions of particular stakeholders, especially data engineers and data scientists. The pipeline can also use embedded machine learning–driven processes to present contextual recommendations and otherwise augment the productivity of data engineers, data scientists, and others involved in building and operationalizing data-intensive apps.

2 Identify priority use cases for a modernized DataOps pipeline

Gaining senior management buy-in to the need for DataOps modernization is only half the battle. Building grassroots support for the initiative depends on how well the modernized DataOps pipeline will help key stakeholders do their jobs.

Priority use cases for a modern DataOps pipeline include business and operational challenges facing the following key stakeholders:

- **Data engineers** can design the orchestration of interdependent DataOps workflows using the pipeline. They can also use it to monitor, manage, accelerate, and improve the flow of ETL, change data capture (CDC), data profiling and cleansing, and other processes that ingest, prepare, and deliver high-quality data to a wide range of business applications.

- **Data analysts** can employ the pipeline to source high-quality data for reports, dashboards, and performance management. They can also rely on the pipeline to help identify the root causes that explain why a particular dashboard is displaying an error or anomalous value.

- **Data stewards** can take advantage of the pipeline to manage data quality in their respective domains. This involves leveraging the pipeline to decrease the frequency and severity of bad data and to catch and resolve these issues before they can impact downstream analytics, decision-support scenarios, and process management scenarios.

- **Data scientists** can work with pipelines to ensure data reliability and quality across algorithms, models, features, and sources, and to detect data changes that signal drift in a model's features and decay in its predictive accuracy over time. When deployed in conjunction with an MLOps pipeline, a modern DataOps pipeline can prove essential to the success of advanced analytics use cases, such as data exploration and discovery, self-service business intelligence, and democratized data science. It can help organizations incorporate additional data sources, structures, formats, and integration or governance workflows to support additional use cases. It can also facilitate greater automation

of these functions while improving the accuracy, consistency, and reproducibility of the resulting applications.

- **Data architects** can use the pipeline to help optimize the costs and maximize the returns on keeping data assets aligned with business objectives. By mapping cross-system dependencies, data architects may use them to predict how changes to linked business processes trigger complex updates across scattered information systems.

- **IT professionals** can utilize a modern DataOps pipeline to improve system reliability and performance, prevent incidents, and improve data system throughput. They may use these tools to declutter data and thereby reduce the cost of migrating it to the cloud. Another use is for comparing the price and performance of alternative data architectures as well as the corresponding migration costs associated with moving to any of them.

3 Align DataOps modernization with strategic cloud data platform implementation

Many organizations are undertaking DataOps modernization in conjunction with larger enterprise data platform migration and modernization projects. Indeed, DataOps pipelines are a key component of many enterprises' plans to combine their data warehouses, data lakes, and/or unified analytics platforms into data lakehouses.

When migrating DataOps pipelines to a target cloud data platform, be sure to consider implementing the following modernization approaches:

- **Separate storage and computing resources for DataOps jobs.** This approach enables rapid scaling of data engineering workloads. Storage and computing elasticity are essential for robust DataOps pipelines, which not only maintain production data but often have many copies of the corresponding source and in-process data sets residing in the non-production environments. It enables more efficient data distribution across the zones, clusters, and nodes that make up the target cloud infrastructure. Separating storage from computing resources can facilitate more intelligent workload redistribution between ETL and ELT approaches. It can also enable data engineering jobs to be more readily provisioned with guaranteed isolation from one another.

- **Process DataOps workloads on a multimodel cloud data platform.** This approach involves managing heterogeneous data in a multimodel, no-copy cloud data platform. Within such a DataOps platform, specialized processing engines are loaded on demand. For any specific DataOps workload, the most suitable processing engines vary in accordance with the relevant data types and formats, the transformations and other functions to be performed on them, and the latency requirements of the downstream platforms and applications that will consume their output. Massive processing of unstructured data may require a horizontally scalable document database, while real-time data may optimally be processed in an in-memory database, machine data in an event store, clickstream data in a key-value database, behavioral data in a graph database, and so on.

- **Migrate DataOps processes incrementally.** This involves starting small and incrementally building out the cloud DataOps pipeline. To mitigate the risks of disruption to source and target systems, move specific data subsets and corresponding ETL code, then incrementally scale up before undertaking the next data/code migration.

- **Automate DataOps change management.** This involves automatically detecting a change in a data source, triggering movement of the data along with any corresponding changes to schemas and formats to downstream systems in the DataOps pipeline. Without strong automation of change data capture (CDC) processes, changes to source data risk making those changes unavailable for a period of time to downstream systems, which can be highly disruptive to applications and processes that depend on that data.

- **Leverage serverless functions for on-demand DataOps processing.** This approach involves on-demand, event-driven, asynchronous, and stateless processing of data engineering workloads. It enables elastic scaling of the density and efficiency of compute, storage, data processing, and other resources. It is fully managed by a data cloud provider so users only pay for the resources when needed, not (as with traditional cloud computing) for always-on apps and servers.

4 Make the necessary investments in enabling infrastructure, tools, and skills for DataOps modernization

DataOps modernization rides on a deepening stack of sophisticated cloud infrastructure and tools. When modernizing the DataOps pipeline, be sure not to skimp on the following investments:

- **Data observability tools** automate unified rollups of end-to-end data provenance, processing, and movement; augment expert assessments of data completeness, consistency, and accuracy; accelerate business responses to compliance, auditing, quality, and other challenges that hinge on data transparency and explainability. These tools also assure enterprises that their high-quality data is always delivered reliably, efficiently, and continuously to all users. They enable DataOps teams to continuously predict, prevent, and resolve incidents in data processing, storage, and other pipeline platforms. They also support configuration, tuning, scaling, monitoring, management, and optimization of the end-to-end DataOps pipeline.

- **Intelligent ETL tools** span hybrid cloud, multicloud, and other complex DataOps pipeline topologies. They are optimized for real-time, low-latency, and continuous processing. These tools can adapt dynamically to changing contexts, workloads, and requirements. They integrate smoothly with MLOps pipelines for more efficient data preparation and training of ML models into intelligent applications, and they

automatically generate real-time contextual recommendations that guide DataOps professionals in the management, optimization, and troubleshooting of workflows and jobs of various degrees of complexity.

- **Data governance and metadata management tools** support versioning of reference data, predictive features, and ML models to reuse in AI, advanced analytics, business intelligence, and other data-driven apps. They also oversee the pool of data that describes the location, provenance, usage, and other attributes of an enterprise's transactional, analytics, and other application data. Modern tools work with technical metadata (which describes the structures, components, and types of data), business metadata (which describes data in user-friendly terms that people with basic tech skills can understand), and operational metadata (which records access to—and usage of—data by individuals, applications, and systems).

- **Data catalogs** maintain searchable inventories of application data and associated ML models, metadata, glossaries, and other semantic attributes. Modern data catalogs support discovery of relevant metadata, scanning of each new data set for sensitive data, automated cleansing of data, suggesting missing lineage between data sets, and tagging data for security and curation purposes.

- **Data lineage and impact analysis tools** identify how specific data items originated, how extensively they have been transformed and cleansed, and how widely they have been distributed. Modern data lineage tools can identify why a particular dashboard is displaying an apparently anomalous value, illustrate the flow of ETL processes that populate data into business applications, or predict how changes to linked business processes trigger complex updates across scattered information systems.

- **Master data management tools** maintain consistent reference definitions and glossaries of business entities (e.g., customer or product) and data about them across multiple IT systems. For instance, many companies want a 360-degree view of each customer because it helps the organization retain and grow that customer.

- **Data stewardship and curation tools** enable efficient discovery, profiling, matching, merging, correcting, tagging, and enhancing of data for delivery to enterprise data warehouses and other repositories.

- **Source-control repositories** provide developers visibility into all the models, code, APIs, containers, virtual machines, business rules, and other multicloud application pipeline artifacts. They also serve as the hub for collaboration, reuse, and sharing of all DataOps pipeline artifacts.

5 Bring simplicity into the modernization of the DataOps pipeline

Implementation of a modernized DataOps pipeline should span an enterprise's many business units and application domains in order to deliver the greatest value to the business.

Modernizing a DataOps pipeline can be a project with many moving parts. To deliver on the

objectives of this project while ensuring that its complexities don't become unwieldy to implement or manage, enterprises need to simplify their modernization efforts.

Simplification is the heart of many DataOps modernization initiatives because many such efforts focus on converging heretofore siloed data engineering and governance workflows. When designing a unified DataOps pipeline, enterprise IT professionals should keep these simplicity pillars uppermost in mind:

- **Scalability:** When modernizing DataOps pipelines, your team's focus on scalability from the project's start will help you support growing workloads smoothly and predictably. This will help you avoid rethinking, reconfiguring, rebalancing, or otherwise adding complexity to the underlying processing resources as workloads expand.

- **Modularity:** By composing a general-purpose DataOps pipeline as a set of modular processing zones, enterprise IT professionals are laying down a blueprint for flexible evolution and composable programmability of that utility in line with changing requirements without adding unnecessary architectural complexities.

- **Observability:** Simplifying the administration of DataOps workloads requires the proverbial "single pane of glass" to monitor the status of all processes, jobs, and functions across the entire pipeline. Monitoring the health of the end-to-end pipeline enables data engineers to drive continuous delivery of reliable data analytics to downstream applications. This requires visibility from the initial ingestion through every intermediate processing step, all the way to delivery of refined data, analytics, and ML models.

- **Automation:** Leveraging AI to proactively provision resources to DataOps workloads and respond to technical, workload, and performance issues, automation enables organizations to scale *up* their data pipelines while simultaneously simplifying and scaling *down* the human effort needed to manage them 24/7.

More broadly, a converged DataOps pipeline can be the target infrastructure for the convergence of multiple MLOps silos. A modernized DataOps pipeline might be designed to ingest and prepare the data used to model, train, and optimize machine learning, deep learning, and other AI models. When unification of DataOps and MLOps pipelines is an objective, various implementation approaches should be explored.

A converged DataOps/MLOps pipeline may be executed through physical and/or virtual consolidation in various scenarios:

- If the current DataOps environment is robust and can handle MLOps workloads efficiently, consider shifting these away from siloed MLOps platforms

- If the enterprise has made a considerable investment in a cloud platform with strong metadata management, semantic integration, and business glossary capabilities, consider physically migrating legacy DataOps and MLOps pipelines to this environment

- If the target cloud platform offers strong data virtualization capabilities, logical unification of legacy DataOps and MLOps pipelines may be the preferred option

6 Restructure the DataOps pipeline in the process of modernizing it

Modernization can be a prime opportunity to restructure how a DataOps pipeline is managed within the enterprise's IT, data analytics, and application development processes.

When aligned with technical enhancements, restructuring initiatives such as the following can make a DataOps pipeline more efficient, effective, and manageable:

- **Staffing:** Centralize DataOps pipeline design, operations, monitoring, and management. Designate an enterprise architect to spearhead the design and implementation of the modernized DataOps pipeline. Upgrade technical staff skills and certifications to support the converged, modernized DataOps pipeline. Rethink DataOps job descriptions to reflect the increasingly automated nature of many tasks and the growing need for higher-level stewardship and governance positions to maintain data quality and pipeline health metrics.
- **Standardization:** Institute standardized, documented cross-enterprise DataOps processes, roles, metrics, and observability dashboards. Adopt standard DataOps APIs, platforms, and formats. Ensure that DataOps workflows and role definitions align with standard DevOps practices.
- **Productivity:** Automate as many DataOps functions as possible. Provide all stakeholders with standardized self-service, visual, no-code, collaborative, and task-focused DataOps tools. Augment staff productivity with ML-enabled recommenders and digital assistants that surface relevant, task-oriented DataOps guidance.
- **Metrics:** Provide all DataOps stakeholders with real-time dashboards, visualizations, and other tools to help track, manage, and report the pipeline metrics that matter most. Chief among these are metrics that track the quality, relevance, and trustworthiness of the data processed within and delivered by the pipeline, as well as those that focus on the utilization, latency, reliability, and availability of the pipeline itself.

Concluding thoughts

Enterprises require a modern set of platforms, tools, skills, and techniques for operationalizing data and analytics. A modern DataOps pipeline ensures that these assets are always fit for purpose and that the workflows needed to achieve this outcome are always scalable, repeatable, high-performance, and transparent.

To evolve their organizations' DataOps pipelines to address fresh business and operational challenges, data management and analytics professionals should:

- Predicate the justification for pipeline modernization on the potential for boosting the speed, efficiency, and effectiveness of the workflow that operationalizes high-quality data and analytics into production applications

- Build stakeholder support for and adoption of the modernized pipeline by showing how well it helps key stakeholders—such as data engineers, data scientists, and data analysts—to do their jobs

- Coordinate implementation of the modernized DataOps pipeline with strategic enterprise cloud data platform migrations, such as converging data warehouses, data lakes, and unified analytics platforms into data lakehouses

- Supplement the rollout of the modernized DataOps pipeline with a deep stack of sophisticated cloud infrastructure for observability, ETL, data and model governance, data quality and profiling, data cataloging and metadata management, and other critical functions

- Architect the DataOps pipeline for simplicity—hence low cost of ownership and administration—by building comprehensive modularity, scalability, observability, and automation into its overall design

- Restructure how the DataOps pipeline is managed as modernization enhancements proceed, making sure to address necessary adjustments to staffing, standards, collaboration tools, and observability metrics

Considering the deep stack of infrastructure and tools necessary for full-fledged DataOps modernization, enterprises should consider acquiring these capabilities as composable stacks from commercial software vendors or DataOps SaaS providers. Bespoke development of these capabilities can be expensive, complex, and error-prone, so it makes sense to buy it all through packaged commercial and/or open source offerings that can be tweaked, extended, or otherwise customized to your enterprise's specific needs.

About our sponsor

gathr.ai

Gathr.ai

Gathr.ai is a next-gen, cloud-native, fully managed, no-code data pipeline platform. It's an all-in-one platform for all your data integration and engineering needs—batch and streaming ingestion, CDC, ETL, ELT, data preparation, machine learning, and analytics. The platform is built on open source, purpose-built data engineering frameworks. It brings unmatched speed, performance, and flexibility required to handle all types of data and analytics approaches in ways that traditional ETL tools cannot. With Gathr's visual drag-and-drop interface, native integration for all popular data sources and destinations, an exhaustive set of prebuilt operators, and a rich pipeline template gallery, anyone can build and deploy data pipelines, quickly and easily.

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, assessments, inquiry services, and topical conferences as well as strategic planning services to user and vendor organizations.

About TDWI Checklist Reports

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, analytics, or a related data management discipline. Companies may use this overview to get organized before beginning a project or to identify goals and areas of improvement for current projects.

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