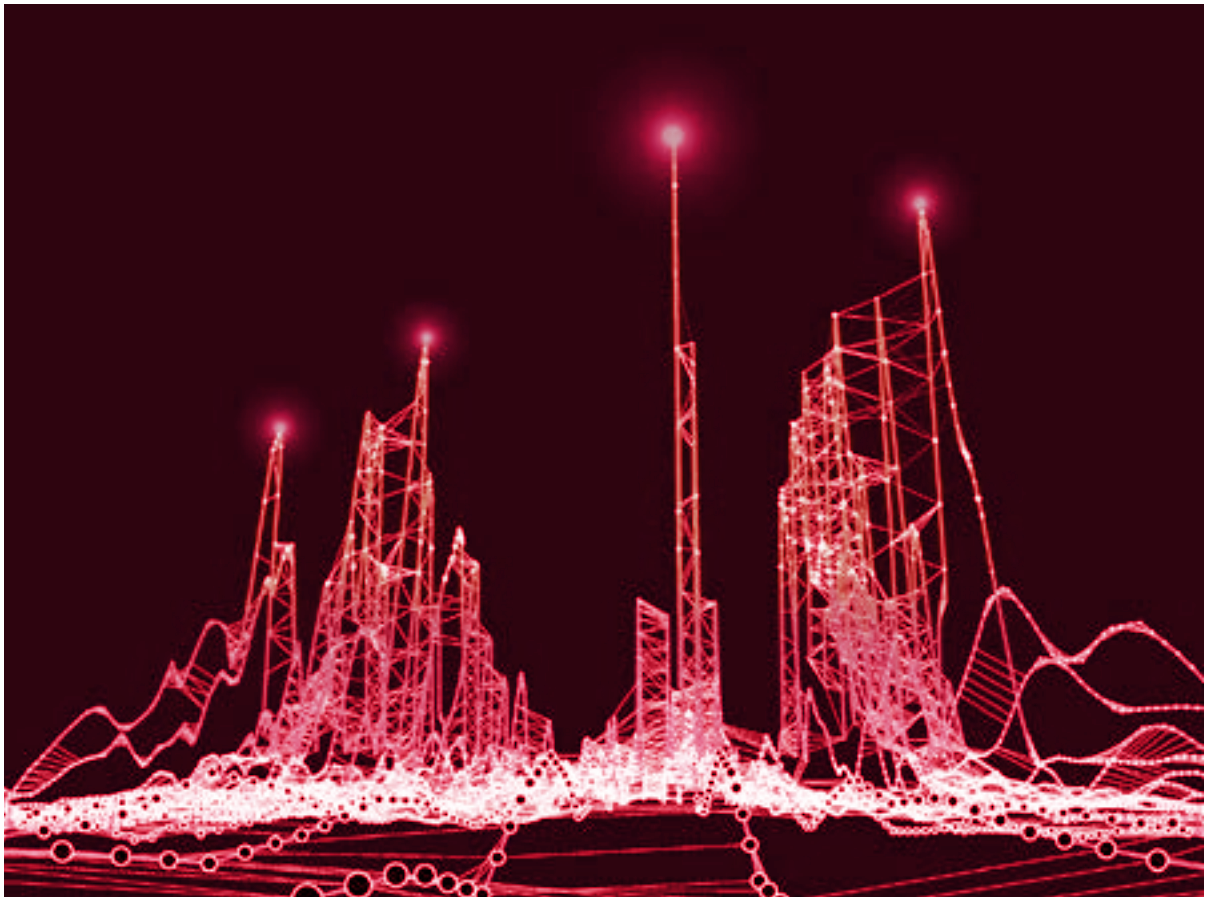
playbook

Modernizing Data Pipelines

By David Stodder



Sponsored by:

HITACHI
Inspire the Next

tdwi TRANSFORMING DATA WITH INTELLIGENCE™

Introduction

In a building, plumbing is part of the behind-the-walls infrastructure, functioning best when unseen and unheard. A data infrastructure similarly depends on pipelines to deliver data seamlessly to the right users, algorithms, and applications at the right time and in the right condition. A building's plumbing subsystems can be complex (perhaps more so than some would wish), but pipelines in a large data infrastructure take complexity to an entirely different level.

Organizations want to tap hundreds, if not thousands, of data sources, with new sources becoming available all the time. Diverse use cases drive new data curation and processing requirements. Organizations need to control access to sensitive data and ensure adherence to rules and regulations governing data use and sharing. Unlike a building's plumbing that needs only occasional maintenance unless there is a disaster or major remodeling, data pipelines require continuous modernization so they can take advantage of the latest tools and data processing platforms. Otherwise, data pipelines become difficult and costly legacy problems that thwart data-driven objectives.

Data pipelines are essential to delivering data for daily business decisions, analytics (including artificial intelligence and machine learning), and data-driven applications. This TDWI Playbook focuses on solving data pipeline challenges so you can grow your organization's data infrastructure with confidence and future-proof it to handle new data and use cases. First of all, what exactly is a data pipeline?

Definitions vary because data pipelines vary. Broadly, a data pipeline is a sequence of steps that move or replicate data from its original sources ultimately to the users' data platform, such as a data warehouse or data lake. Organizations frequently set up additional data pipelines between data lakes and secondary target destinations such as analytics databases or data marts. In between sourcing data and moving it to targets, data pipelines can include curation steps for profiling, cleansing, transforming, and enriching data as needed to serve business intelligence (BI) reports, dashboards, analytics,

artificial intelligence and machine learning (AI/ML) development, and applications.

Data sources for pipelines are varied. These can include raw, unstructured data such as log files that have not been adjusted, cleansed, or manipulated. Many pipelines deliver semistructured data that uses HTML, JSON, or XML formats; others extract structured transactional data from multiple business applications and e-commerce systems to load into staging areas for transformation or directly into target data platforms.

The fact that both data sources and target data platforms are evolving presents additional challenges. For example, organizations are digitally transforming applications and migrating from on-premises data systems to cloud-based data lakes and data warehouses that use public cloud services. However, most organizations in TDWI research have data sources and targets both on premises and in the public cloud (aka hybrid cloud). Data pipeline infrastructures need to address hybrid multicloud environments.

Data pipeline development needs to be systematic and scalable. Traditionally, data engineering teams are responsible for data pipelines. However, because most data engineers focus on technology issues, many lack a "data as a product" perspective, which is central to modern frameworks such as the data mesh and DataOps. Data as a product is about provisioning data more precisely to fit the business purposes of users and applications.

Data engineering practices tend to be ad hoc and have lagged behind software engineering practices that have evolved, with DevOps principles, to be more agile and automated. To meet growing data-driven business demands, organizations need similar frameworks for data that free them from spending too much time stitching data together using ad hoc pipeline processes. They need a systematic approach such as DataOps that makes it easier to corral the explosion

in data volumes, sources, types, and velocity. Data-driven business demands require treating data as a product that enables "customers" (data consumers and other stakeholders) to realize faster business value through better decisions, insights, and actions.

Pipelines need scalable technologies for data collection and ingestion. Once sources are identified and located, collection and ingestion processing steps traditionally occur in batch mode—for example using extract, transform, and load (ETL) tools. An increasing number of organizations today reorder the sequence to ELT, loading the data first into a target data platform and running transformations there using powerful, often cloud-based, in-database processing. Modern data pipeline practices and technologies must support both ETL and ELT processing.

However, to feed operational reporting and data science operations that use predictive models and AI/ML algorithms, many organizations are shifting from batch mode to ingesting data via continuous, real-time data streaming. This is particularly interesting for enterprises using data from industrial data sources such as operational technology (OT) sensors and combining it with IT data. Another alternative to batch processing is incremental processing using change data capture (CDC) tools, in which organizations capture and send just the changed data to a data warehouse, data mart, specialized data lake zone, or operational reporting system.

TDWI research suggests that organizations need to modernize tools and practices to handle both batch and streaming, event-driven processing. More than two-thirds of research participants (68 percent) say their developers face challenges processing streaming data and CDC updates.[1]

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

Important additional attributes of modern data pipeline systems include the following (some of which we will discuss later in this TDWI Playbook):

- Templates for code generation to improve reuse and avoid unnecessary manual work; support for both manual and event-based data ingestion.
- Decoupled data access to reduce dependencies and enable users to interact with data independently, without disrupting other workloads.
- Continuous management, monitoring, and testing capabilities for code, data, models, and infrastructure versioning.
- Data observability that offers automated capture of metrics and usage information to improve pipeline performance.
- Continuous integration and continuous delivery (CI/CD) to systematize practices in data pipeline development life cycles for scaling up delivery and deployment. CI/CD practices, critical to DataOps, promote automation, code sharing, and integration.
- Flexibility to work with different data targets including data warehouses, data lakes, unified data lakehouses, data hubs, and distributed environments such as a data fabric.
- End-to-end data orchestration as an alternative to the chaos of ad hoc data pipeline development.

Top Data Pipeline Challenges

In TDWI research, over one-third of organizations surveyed (36 percent) need some improvement with their data pipeline development and operationalization and 22 percent are looking for

Which of the following steps are most important to your organization for improving data pipelines and data preparation? Select up to five.

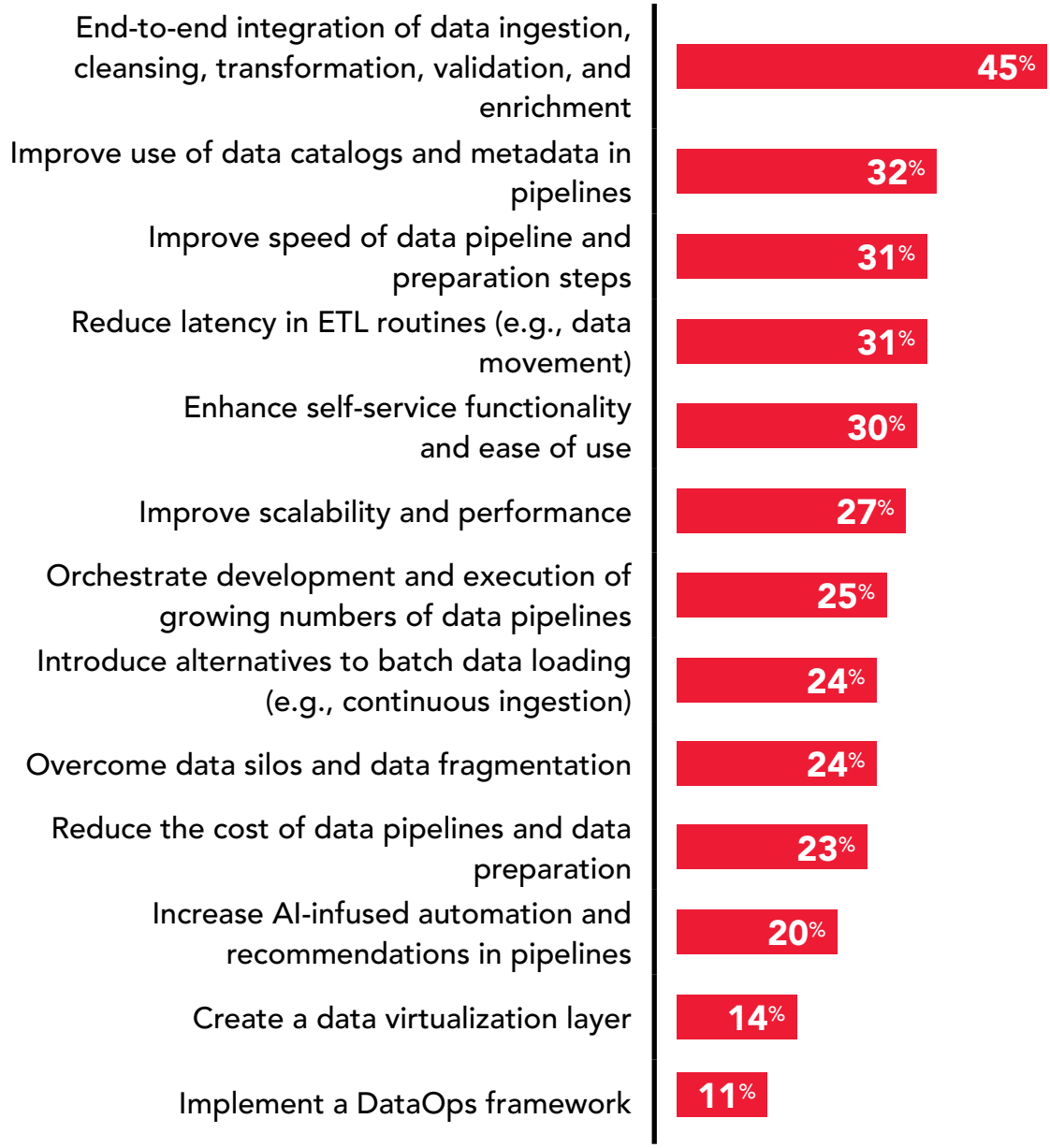


Figure 1. Based on answers from 364 respondents. Source: Q4 2021 *TDWI Best Practices Report: Modernizing Data and Information Integration for Business Innovation.*

a major upgrade to technologies and practices.[2] Given the growing number and complexity of data pipelines, an inability to manage and orchestrate processes is a major concern. Figure 1 shows that gaining end-to-end integration of data ingestion, cleansing, transformation, validation, and enrichment is the top priority for survey respondents.

The second-highest percentage of organizations in Figure 1 note the importance of data intelligence to modernization. Data intelligence is knowledge about the data, its lineage, and its relationship to business entities of interest. Most organizations manage data intelligence in an enterprise data catalog or another metadata repository. Along with definitions, data intelligence can include information about the data's quality, reliability, and availability, all of which is essential to improving the value of data as a product.

Developers, data engineers, and analysts can apply data intelligence during data pipeline processes to locate and access trusted data faster and more completely. In Figure 1, 32 percent want to improve the use of data catalogs and metadata in pipelines. Automating integration between data pipeline processes and data catalogs is fundamental to modernization.

Third in importance in Figure 1 is improving speed in data pipeline and preparation steps. With organizations placing a high priority on data-driven decision-making, speed to insight is a competitive advantage. Monitoring is critical for troubleshooting reasons for substandard performance, such as bottlenecks in data flows, incomplete jobs, and scalability constraints that affect processing speed. Problems with integrating fragmented data across data silos cause delays and can increase latency. About one-quarter of organizations surveyed by TDWI say that overcoming silos is important to improving data pipelines and data preparation.

We will discuss these challenges further as we examine business drivers and define critical steps for getting started with modernization.

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

Business Drivers for Modernizing Data Pipelines

Especially in times of rapid change, organizations need data pipelines that accelerate critical data to users so they can make informed decisions and gain business insights. Organizations need to tap internal and external sources to learn about customers in order to engage with them effectively across multiple channels and in real time. Data needs to flow into analytics to uncover previously hidden opportunities, mitigate risks, and enable proactive steps for better outcomes.

TDWI finds that the following three business drivers commonly motivate strategies to modernize data pipelines:

Data democratization and self-service data interaction. Organizations want to empower a greater variety and number of business users to use self-service capabilities to discover, blend, visualize, and analyze data. In TDWI research, 41 percent of organizations are using data pipelines for BI reporting, dashboards, and business analytics.[3] However, business users struggle to develop pipelines on their own, often due to the technically skilled manual work required. If business users rely on data engineers to develop data pipelines and prepare data for them, it takes them longer to reach valuable insights and they have less time to apply those data insights to their work.

Data science and AI/ML. Data science projects can bring competitive differentiation and innovative process changes that increase efficiency. Data scientists, engineers, and analysts often work with diverse, high-volume data and raw data streams. Citizen data scientists—that is, business users who are data-savvy and want self-service analytics capabilities—are increasing in number. The combination of data democratization and growth in data science projects puts pressure on organizations to orchestrate diverse data pipelines.

[3] Research about use cases for data pipelines is from the 2022 *TDWI Best Practices Report: Maximizing Business Value with Data Platforms, Data Integration, and Data Management*, online at tdwi.org/bpreports.

Modern data-driven applications feature embedded analytics and use AI/ML to guide automated decisions. They require reliable data pipelines to support interactions with thousands, if not millions, of employees, partners, and customers around the globe. In TDWI research, 35 percent of organizations are using data pipelines for applications and 21 percent for customer or partner engagement or data sharing. Scalability, speed, quality, and resilience are essential.

Treating data as a product. With its heightened role in nearly all products and services, many organizations are changing how they think about data. Cutting-edge organizations view data as a product. Rather than a separate, IT-managed resource, leaders see data as a product with business value generated by the deployment of data-rich products and services.

Many leaders who see data as a product also view data as code; a data warehouse, for example, is a code base that serves internal data consumers. "Data as a product" strategies demand that, just like applications, data should have service-level agreements (SLAs) that set out expected levels of quality, reliability, and performance.

Critical Plays for Getting Started

In this section, we discuss six strategies for getting started with data pipeline modernization and responding to business drivers.

ADOPT A DATA-AS-A-PRODUCT MINDSET

Successful modernization requires you to take a different view of data: that is, as a product that has potential value. In this view, pipelines are more than just plumbing that flows data from one location to another; pipelines have an active role in building the value of data as a product. Developing SLAs based on business value rather than just the delivery of data will help you gain a business perspective on the success of data pipeline processes for data quality, reliability, availability, interoperability, and reproducibility.

Assigning responsibility is an important step. Some organizations hire data product managers who have experience in product or software management rather than IT data management. Chief data officers (CDOs) also lead change toward viewing data as a product. To ensure that data is meeting SLAs, achieving full potential as a product, and satisfying data consumers, pipeline processes need capabilities for testing, monitoring, and documentation.

IMPLEMENT DATAOPS PRACTICES AND TECHNOLOGIES

DataOps practices and technologies can enable your organization to manage and orchestrate data pipelines holistically within a defined framework. DataOps uses the notion of a data supply chain to integrate and operationalize data pipelines faster. It encourages team collaboration and improves focus on increasing the product value of data.

DataOps borrows from agile and DevOps methodologies to give organizations a framework for eliminating delays and inefficiencies through collaborative management, orchestration, and continuous improvement. DataOps can help your organization deal with change; traditional IT practices are often too rigid to handle changing data requirements. Beyond technology modernization, DataOps offers a framework for cultural change, such as the change to viewing data as a product. Frameworks support applying software development disciplines to data pipeline development and deployment.

The remainder of these recommendations are each important on their own, but they are also attributes of DataOps practices.

APPLY DATA INTELLIGENCE TO CLOSE GAPS BETWEEN BUSINESS KNOWLEDGE AND PIPELINES

Data intelligence is critical for enabling your data product managers to answer questions about your data's origin, lineage, and purpose. Accessible, accurate, and up-to-date intelligence, often managed in a data catalog, shortens the time it takes people to find, prepare, and use trusted data in pipelines. Modern data catalogs are automating previously manual processes and employing AI/ML for better accuracy and scalability.

Most critically, data intelligence is important for aligning knowledge of your business use cases, defined in business terms, with knowledge about your data. Using data intelligence based on accurate metadata, master data, and business glossaries, your organization can improve the design and implementation of data pipelines to meet business needs. Documentation is essential to making it easier to adjust pipelines when business context changes. Data intelligence enables you to determine the appropriate data quality, transformation, and other preparation steps needed to enrich your data's value. It can even help you discover sensitive information such as personally identifiable information (PII) and protect it using data privacy pipelines.

CREATE THE ARCHITECTURE AND SELECT TECHNOLOGY BEFORE BUILDING AND TESTING PIPELINES

Data demands can be loud and insistent. However, to avoid the problems of ad hoc data pipeline development and deployment, your organization needs to see pipeline building and testing as a stage that comes *after* developing an overall architecture and selecting the appropriate technologies. Ad hoc development leads to inconsistent and redundant pipelines, poor data quality, and little reuse.

With the architecture and DataOps framework in place, you will have the big picture needed to deploy the right technologies required by your use cases and workloads, such as streaming technology for real-time OT data. Your organization can then develop, test, and deploy data pipelines within your architecture and framework, avoiding proliferation of disparate pipelines that become hard to manage, govern, and maintain.

Some organizations use a hub-and-spoke model, making DataOps the central hub with monitoring capabilities to oversee pipeline development and testing "spokes." Pipelines are then coordinated, but you have flexibility to design and develop new pipelines to meet specific business needs.

ORCHESTRATE AND AUTOMATE FOR SCALABLE, END-TO-END DATA PIPELINE PROCESSES

As data pipelines tap more sources and jobs increase in complexity, orchestration and management become necessary. When your organization builds more jobs, a lack of automation can add operational overhead and increase the potential for human error. Also adding complexity is that most organizations have multiple teams building their own data pipelines in a self-service fashion to produce data products, often without collaboration. It becomes difficult to coordinate, sequence, and operationalize jobs and ensure each one has appropriate resources.

DataOps offers a collaborative framework for implementing tools that can automate jobs and improve end-to-end integration of data pipeline processes. The DataOps principle of continuous integration and continuous deployment (CI/CD) facilitates tighter integration between design and testing. For CI/CD to work, organizations need automated tools that give design teams immediate validation of testing results and feedback for determining the cause of failures, including unexpected changes to the data (i.e., data shift). Automation can accelerate pre- and post-deployment testing. Some automated tools solve problems automatically through self-healing capabilities.

Workflow management and orchestration are critical for avoiding delays and errors that affect business operations and decisions. Tools help; for example, your teams can choose from orchestrator solutions in the marketplace to fit their pipeline workload requirements. Some orchestrators support pre-set operations for performing actions; others have utilities for customizing operations to fit job requirements. You should use modern workflow and orchestration tools to increase speed, scale, and standardization.

DataOps supports end-to-end design and development. Stakeholders can collaborate to ensure understanding of entire data life cycles from origination to use. Tools offering holistic visibility into data pipelines show dependencies between pipelines that may affect performance.

Holistic visibility into end-to-end processes makes it easier for you to choose the right data ingestion and transformation tools.

IMPROVE MONITORING AND DATA OBSERVABILITY

Continuous monitoring is essential as your data environment grows in size, complexity, and impact on business decisions. Monitoring is a core element of DataOps, especially for troubleshooting so you can fix problems and debug coding errors as soon as possible. However, traditional predefined monitoring rules and metrics are proving to be too rigid and limited in data environments that exhibit constant change and expansion.

Many organizations today are supplanting traditional monitoring with *data observability*. Data observability incorporates (or functions alongside) traditional monitoring, but the broader focus is to offer actionable visibility into the health of the data throughout life cycles, not just standard monitoring metrics. Solutions use analytics, including AI/ML, to alert you to changes in the lineage, quality, and availability of data and make it easier to discover unknown business or system events affecting the data. Data observability can also address data governance requirements.

Your organization needs modern tools that enable data engineers to monitor source data as it comes through numerous pipelines. For engineers and users, monitoring can help validate the data and determine how to transform it for user or application requirements.

Monitoring and observability are helpful for data pipelines used in predictive modeling and AI/ML development. Models and algorithms can decay over time, becoming less accurate and trustworthy. A major reason is data drift, where data structures, semantics, and infrastructure change over time, often unexpectedly and without documentation. Monitoring and observability tools can track data drift, making it easier and faster to address problems by taking remediation steps or selecting new data sources.

Conclusion: Modernize and Unify to Maximize Value

Modernization involves adopting new tools that improve automation, including through embedded use of AI/ML to increase data scalability, accuracy, and provisioning. However, organizations are often challenged by the sheer number of tools set up for specific data silos and single, point-to-point data integration. Hybrid multicloud data environments make distributed silos and tools even more difficult and costly to manage.

Organizations should evaluate opportunities to unify tools into a single platform in the cloud. This would give you the opportunity to evaluate your tools, eliminating those no longer delivering value. You can then focus modernization on value-driving use cases and workloads. Your organization could move beyond the complexity of multiple architecture stacks and tools across a hybrid multicloud environment to a single, unified platform.

DataOps frameworks enable organizations to gain a holistic view of data pipeline processes to see the effectiveness of current tools and platforms and focus modernization efforts. This report has discussed how your organization can use DataOps to improve collaboration and address change management, including the shift to modern tools for designing, developing, orchestrating, and deploying data pipelines.

A critical attribute of DataOps is the ability to envision data's importance as a product that is critical to achieving business outcomes. Within this context, you can more clearly evaluate the impact of taking steps to modernize data pipeline processes, streamline end-to-end integration, and scale up to meet data-driven business demands.

Accelerate Business-Ready Data with Lumada and Pentaho

(Content supplied by Hitachi Vantara)

Today's data-onboarding projects require more than just connecting to or loading data from a variety of data sources across on premises and the cloud. Rather, they involve managing a changing array of data sources from a variety of formats, establishing repeatable processes at scale, and maintaining control and governance across all enterprise data. Whether an organization is implementing an ongoing process of onboarding hundreds of data sources into a cloud data lake or enabling business users to prepare diverse data without IT assistance, integration of data still experiences major obstacles:

- Overloaded IT resources
- Putting project deadlines at risk
- Opportunity costs of higher-value tasks
- Repetitive manual design
- Time-consuming development
- Manual error risks

Hitachi Vantara's data integration and analytics software is powered by Pentaho technology. Part of the Lumada DataOps portfolio, Pentaho simplifies managing enormous data volumes with increased variety and velocity entering organizations today. By allowing data preparation from any source and automating your data pipeline, Pentaho accelerates getting data in the hands of your business user for a quicker time to analytics value. This software also delivers business analytics to end users faster with visual tools that reduce time and complexity—without writing SQL or coding in Java or Python. Organizations immediately gain real value from their data, from sources such as files, relational databases, Hadoop, and more, which are in the cloud or on premises.

Leading companies depend on the data management expertise of Hitachi to help them integrate, manage, and govern their data and applications. Pentaho has an adaptive big data execution layer that allows you to plug into popular big data stores with flexibility and insulation from change. Data can be accessed once, then processed, combined, and consumed anywhere.

This adaptive big data layer includes plug-ins for popular cloud data platforms such as AWS, GCP, and Azure, object stores including Hitachi Content Platform (HCP), Hadoop distributions such as Cloudera and HPE Ezmeral Data Fabric, and NoSQL databases including MongoDB and Cassandra.

With broad connectivity to any data type and high-performance Spark and MapReduce execution, Pentaho technology simplifies and speeds the process of integrating existing databases with new sources of data. Pentaho Data Integration's graphical designer includes:

- An intuitive, drag-and-drop designer to simplify the creation of analytics data pipelines
- A rich library of prebuilt components to access, prepare, and blend data from relational sources, big data stores on premises or in the cloud, enterprise applications, and more
- The ability to spot check data in flight with immediate access to analytics, including charts, visualizations, and reporting, from any data prep step
- Powerful orchestration capabilities to coordinate and combine transformations, including notifications and alerts
- An integrated enterprise scheduler for coordinating workflows and a debugger for testing and tuning job execution

Pentaho Data Integration speeds performance, reduces the complexity of integrating big data sources, and provides:

- Code-free data transformation design that empowers 15 times faster productivity versus hand coding and executes in-cluster for high performance
- A template-based approach to rapidly onboard data sources into Hadoop via metadata injection feature set
- The ability to seamlessly switch between execution engines, such as Spark and the Pentaho native engine, to fit data volume and transformation complexity
- Support for advanced analytics models from R, Python, Scala, and Weka to operationalize predictive intelligence while reducing data prep time

For years, companies have independently leveraged Pentaho for their on-premises and cloud data management. They have invested in data pipelines, quality, and governance policies enhanced by Hitachi Vantara solutions. We continue to bring new technology innovations and rich functionality to help accelerate cloud migration and modernization projects while reducing cost, time, and risk.

Today, data-driven organizations need access to trusted data in real time to innovate at scale and achieve their business outcomes. Deliver business value by using trusted data for decisions to make a difference in your business. Lumada DataOps powered by Pentaho is an end-to-end platform to integrate all your structured and unstructured data, understand its meaning using AI, and deliver business insights with advanced analytics. With numerous data silos across edge-to-cloud, efforts to standardize on one technology are difficult; customers need a practical way to deliver data to the business. Lumada offers self-service data discovery, preparation, and access for business users while having the appropriate governance controls for the IT teams.

Experience the Power of Pentaho:
https://www.hitachivantara.com/en-us/products/lumada-dataops/data-integration-analytics/download-pentaho.html.

About Our Sponsor

HITACHI
Inspire the Next

Hitachi Vantara, a wholly-owned subsidiary of Hitachi Ltd., delivers the intelligent data platforms, infrastructure systems, and digital expertise that supports more than 80 percent of the *Fortune* 100.

To learn how Hitachi Vantara turns businesses from data-rich to data-driven through agile digital processes, products, and experiences, visit hitachivantara.com.

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, assessments, 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