CHECKLIST 2022

Using Data Observability to Boost Quality and Efficiency Throughout the Data Pipeline: Five Best Practices

By James Kobielus

acceldata

tdwi TRANSFORMING DATA WITH INTELLIGENCE™

Using Data Observability to Boost Quality and Efficiency Throughout the Data Pipeline: Five Best Practices

By James Kobielus

Operational data pipelines are the unsung workhorses of business success. Enterprises must keep constant watch over their data and the pipelines through which it is ingested, prepared, and otherwise made fit for operational use.

Data observability tools enable organizations to continuously track, correlate, assess, predict, manage, and optimize the health of their data supply chain. Observability is an essential tool for monitoring the metrics such as quality, availability, reliability, efficiency, and performance that businesses apply to the operational data pipeline.

This TDWI Checklist discusses five best practices for using observability tools to monitor, manage, and optimize operational data pipelines. The Checklist provides strategic guidance for C-level executives to articulate the value of data observability to the business. It also provides DataOps professionals with practical steps for implementing, managing, and optimizing data observability in line with their companies' evolving IT, cloud computing, and data management strategies.

Five best practices for using observability tools:

1. Make the strategic case for data observability
2. Identify the chief business metrics of data observability
3. Deliver actionable observability to every data stakeholder
4. Instrument the enterprise data infrastructure for comprehensive observability
5. Use intelligent observability to augment data pipeline professionals' productivity

1 Make the strategic case for data observability

Enterprise data is often a complex entity with a significant backstory. Within a data processing pipeline, a particular record may be stored, merged, cleansed, and otherwise processed at various points and in many ways before being delivered to its intended use.

Data observability tools illuminate the story of how this precious asset originated, how extensively it has been processed, how and by whom it's been used, and how widely it's been distributed. In conjunction with data catalogs, business glossaries, metadata management, and other capabilities, data observability tools can help business and IT stakeholders do their jobs more effectively.

Making the business case for enterprise-wide data observability is straightforward. These tools provide the following functions:

- Automate unified roll-ups 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
- Provide assurance that high-quality enterprise data is delivered reliably, efficiently, and continuously to all users
- Help organizations understand the state of their data and the systems that process, store, and manage that data at any point in time, regardless of data's source, technology, or scale
- Enable data teams to continuously predict, prevent, and resolve incidents in data processing, storage, and other pipeline platforms
- Support configuration, tuning, scaling, monitoring, management, and optimization of the end-to-end data pipeline

2 Identify the chief business metrics of data observability

Data observability tools provide real-time dashboards, visualizations, and other formats to help businesses track and manage the metrics that matter most.

Chief among these observability metrics are the following attributes associated with the quality, relevance, and trustworthiness of the data that is processed within and delivered from the pipeline:

- **Provenance**—where did the data originate and how has it been handled and moved from point of origin to the present moment?
- **Correctness**—has the data been kept consistently accurate and updated at every point in time for every use?
- **Completeness**—has the data been aggregated, packaged, and delivered along with all relevant content at every point in time for every use?
- **Consistency**—have all copies of the same data been kept in sync at every point in time for every use?
- **Compliance**—does the format, content, handling, and management of data meet all relevant standards, mandates, and requirements?

- **Structure**—to what extent have the schema, distribution, relationships, and metadata of the data changed over time?
- **Clarity**—are the meanings, definitions, and linkages of data transparent and explainable to every user for every use?
- **Impact**—how readily can the consequences of modifications to the content, format, and handling of data be discerned in downstream uses?

Just as important are the following operational metrics pertinent to the pipeline itself:

- **Utilization**—does the data pipeline have the bandwidth, memory, processing, storage, and other resources needed to handle expected workloads while meeting service-level requirements? To what extent are these resources often saturated or used to the maximum?
- **Latency**—how long does it take to deliver various workloads and classes of service over the data pipeline? What is the maximum duration that an item of data must await processing?
- **Reliability**—what is the incidence of failures, errors, and glitches in applications and infrastructure components in the end-to-end data pipeline? To what extent does the pipeline consistently and proactively detect and resolve these problems before they can impact the business?
- **Availability**—what is the frequency of unplanned downtime in every segment of the data pipeline and to what extent does the pipeline adaptively maintain continuity of service through automated rerouting, backup/restore, and other techniques?

3 Deliver actionable observability to every data stakeholder

Responsibility for data and its business uses is woven throughout different teams within every organization. Consequently, everybody in the business needs some level of observability into the data regarding how it's been ingested, cleansed, transformed, and otherwise processed within the pipeline.

Here are some high-level uses of data observability by various stakeholders:

- IT professionals may rely on data observability tools to help them improve system reliability and performance, prevent incidents, and improve data system throughput. They may use these tools to structure and organize 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.
- Data engineers may use observability tools to visualize the flow of extraction, transformation, and loading processes that populate data into a wide range of business applications.
- Data scientists use observability 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.
- Data architects may use observability to help them optimize the costs and maximize the returns on keeping data assets aligned with business objectives. Also, 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.

- Business intelligence developers may use data observability tools to identify the root causes that explain why a particular dashboard is displaying an error or a seemingly anomalous value.
- Data stewards may use data observability tools to decrease the frequency and severity of data quality issues, catching and resolving these issues before they can impact downstream analytics, decision-support scenarios, and process management scenarios.
- Data analysts may use data observability tools to assess the incidence of errors in BI reports, operational dashboards, and other analytics apps that might contribute to bad business decisions.
- Executive stakeholders, such as the chief data officer or vice president for data and analytics, may rely on the enterprise investment in data observability to cost-effectively align data delivery with business outcomes while mitigating the risks of bad data and unreliable DataOps pipelines.

To provide effective support to any of these stakeholders, data observability tools must perform several core functions:

- Enable real-time monitoring of updates and predictive analysis and forecasting of likely changes over various future timescales
- Help data professionals to profile, roll up, track, predict, and resolve operational data issues before they impact the business
- Correlate anomalies and other impactful events across data sources and processing nodes, infrastructure and metadata platforms, and all segments and layers of the pipeline
- Provide visibility into relationships, dependencies, lineage, metadata, and other key attributes of data's end-to-end life cycle
- Support historical trend analysis of metrics, leveraging the timestamping and logging of metrics to enable diagnosis of root causes of data pipeline and data quality issues over time

4 Instrument the enterprise data infrastructure for comprehensive observability

Enterprise IT professionals require comprehensive observability that spans the diverse range of data pipelines, computing platforms, administrative tasks, and use cases for which they are responsible.

Enterprises should instrument their data infrastructures with observability tools that deliver the following valuable capabilities to the business:

- Make the platform accessible to all authorized stakeholders, consistent with their credentials, permissions, roles, responsibilities, and other attributes
- Integrate observability tools into existing data platforms, cloud computing environments, applications, and other IT infrastructure through open APIs and other standard approaches
- Implement a platform-agnostic observability solution that can support legacy on-premises data platforms as well as emerging cloud data environments

- Use APIs to integrate the observability platform with existing data pipelines, management tools, and stakeholder applications
- Implement an observability platform that can scale and adapt to the enterprise data environment
- Monitor streaming data to enable fast response to data quality issues for near-real-time use cases.
- Configure the data observability platform to automatically generate data quality rules
- Set up configurable alerts to help avoid alert overload and make sure that data stakeholders can catch issues by using fine-grained thresholds and controls

5 Use intelligent observability to augment data pipeline professionals' productivity

Data pipelines have become too vast, complex, and dynamic for humans to monitor and manage with purely old school IT system management approaches. Traditional application performance monitoring solutions generally lack the ability to monitor data flow through these pipelines. Keeping pace with the scale and complexity of today's data pipelines requires a new generation of cloud-based machine learning (ML)-powered observability tools. These solutions automate the following core DataOps functions:

- Validating data as it is ingested into the pipeline
- Learning pipeline traffic patterns
- Correlating pipeline events
- Identifying the root causes of anomalies, faults, errors, and glitches in the pipeline
- Fixing pipeline issues before they impact performance or create unplanned outages, cost overruns, or bad output based on low-quality data
- Providing configurable, fine-grained alerts to ensure that data and IT administrators catch critical issues such as pipeline bottlenecks before they can become business showstoppers
- Optimizing methods for accessing, persisting, transforming, enhancing, and delivering data

In addition to automating pipeline runtime functions, modern data observability tools leverage ML to guide data teams in addressing issues whose remediation cannot be entirely automated. They generate recommendations for data engineers on how best to manage, optimize, and troubleshoot jobs of various degrees of complexity. They may use ML to intelligently scan data assets across the enterprise, automatically add business context metadata, automate data tagging and domain/entity recognition, and populate the data catalog with updates to descriptive metadata about the data being processed in the pipeline.

Concluding thoughts

Intelligent observability tools should be a core feature of every enterprise data pipeline. Modern observability solutions enable data teams to predict, prevent, and resolve quality issues within the data while ensuring the continued efficiency, reliability, and availability of the pipelines through which this critical business asset is processed, managed, and delivered.

As discussed in this Checklist, the key best practices in data observability include making the business case for investments in these tools, identifying metrics to be used in monitoring and managing data pipelines, delivering actionable observability to every data stakeholder, and instrumenting end-to-end data pipelines for comprehensive observability.

Enterprises must also be sure to automate the workloads of DataOps pipeline professionals to the maximum extent possible. Automated observability tools enable organizations to scale up their data pipelines while at the same time scaling down the human effort needed to monitor and manage them 24/7. Be sure to invest in ML-driven solutions that proactively detect and resolve issues with enterprise data while also providing contextual recommendations to help IT personnel manage those complex DataOps tasks that cannot be entirely automated.

About our sponsor

Acceldata's Data Observabilty Cloud helps enterprises transform their data systems from unreliable, hard-to-scale, and expensive to stable, agile, and cost-efficient. It correlates events across data, processing, and pipelines to transform how organizations observe, operate, and optimize enterprise data systems. It delivers deep data observability, spanning metrics, logs, and data quality to improve reliability, accelerate scale, lower costs for real-time AI and analytics workloads, and reduce organizational risk. Acceldata's platform has been embraced by global customers such as Oracle, PubMatic, PhonePe (Walmart), Pratt & Whitney, DBS, and many more.

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