Eckerson Group | White Paper

Top Ten Use Cases for Data Observability

Requirements, Techniques, and Guiding Principles

KEVIN PETRIE

JANUARY 2023

RESEARCH SPONSORED BY

acceldata

About the Author

Kevin Petrie is is Vice President of Research at Eckerson Group, where he manages the research agenda and writes about topics such as data pipelines, cloud data platforms, machine learning, and data observability. For more than 25 years Kevin has deciphered what technology means to practitioners, as an industry analyst, author, instructor, marketer, and services leader. Kevin launched and led a profitable data services team for EMC Pivotal in the Americas and EMEA, and ran field training at the data integration software provider Attunity (now part of Qlik). A noted public speaker and author of two books on data streaming, Kevin frequently shares thought leadership content on **LinkedIn**.

About Eckerson Group

Eckerson Group is a global research and consulting firm that helps organizations get more value from data. Our experts think critically, write clearly, and present persuasively about data analytics. They specialize in data strategy, data architecture, self-service analytics, master data management, data governance, and data science. Organizations rely on us to demystify data and analytics and develop business-driven strategies that harness the power of data. **Learn what Eckerson Group can do for you!**

About This Report

Research for this report comes primarily from numerous briefings with software vendors. This report is sponsored by Acceldata, who has exclusive permission to syndicate its content.

Table of Contents

Executive Summary

Data observability uses techniques adapted from governance and application performance management tools to study the health of modern data environments. This emerging discipline includes data quality observability, which studies the accuracy and timeliness of data in flight or at rest, and data pipeline observability, which studies the performance of data pipelines as well as the infrastructure that support them. Data observability programs and solutions should address these ten use cases across four categories:

- ***Prepare.*** *Infrastructure design, capacity planning, and pipeline design.*
- ***Operate.*** *Performance tuning, data quality, and data drift.*
- ***Adjust.*** *Resource optimization, storage tiering, and migrations.*
- ***Fund.*** *Financial operations (FinOps).*

Eckerson Group recommends that data leaders use the following guiding principles to increase the odds of success with a data observability program:

- ***Let the business prioritize your use cases.*** *Each use case described in this report costs time, effort, and money. Engage business owners as you assemble your use cases and identify which ones offer the right mix of cost, benefit, and risk.*
- ***Foster team collaboration.*** *Data engineers, platform engineer, CloudOps engineers, platform engineers, data analysts, data scientists, and finance managers must learn to communicate and help one another. This might require training, dotted-line reporting structures, and periodic cross-functional meetings.*
- ***Standardize on shared tools.*** *Teams collaborate more effectively when they share common tools. Look for a multi-modal platform for data observability that both IT engineers and analytics teams can use, and that business owners can oversee.*

The What and Why of Observing Data

As enterprises drive data into more of their decisions and analytics into more of their operations, they need to ensure that data tells an accurate story. Conflicting or outdated records can confuse stakeholders and lead to harmful actions. Consider the machine learning model that calls legitimate transactions fraudulent, or the sales manager that pitches a customer without knowing they just complained to the support desk. Situations like these underscore the reality that wrong or incomplete data is a liability rather than an asset. To make data an asset, enterprises need data observability.

This report defines data observability, including its challenges and benefits. Then we explore use cases for preparing, operating, and adjusting data environments, as well as managing the business aspects of analytics projects and applications.

Data Observability Defined

Data observability is an emerging discipline that studies the health of enterprise data environments. It uses techniques adapted from governance and application performance management tools to address modern use cases and environments. Data observability tools apply machine learning to monitor the accuracy and timeliness of data delivery, with a particular focus on cloud environments. This helps optimize data delivery across distributed architectures for both analytics and operations.

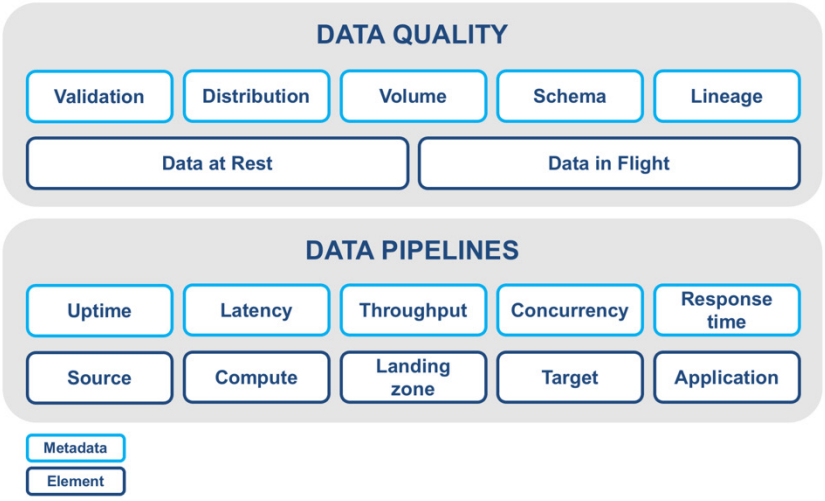
Data observability is an emerging discipline that studies the health of enterprise data environments.

Data observability includes:

- **Data quality observability**, which studies the quality and timeliness of data. It observes data in flight or at rest, for example, by validating sample values and studying metadata that measures value distributions and data volumes, schema, and lineage.
- **Data pipeline observability**, which studies the quality and performance of data pipelines, including the infrastructure that supports them. It observes elements such as data sources, compute clusters, landing zones, targets, and applications. It does this by studying metadata that measures uptime, latency, throughput, concurrency, and response time.

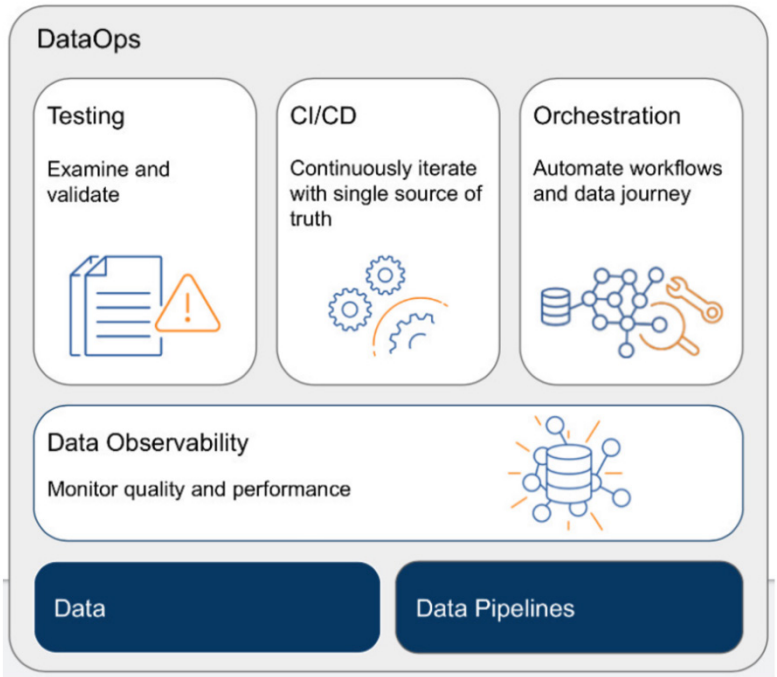
Figure 1 describes the elements and metadata that contribute to data observability.

Figure 1. Data Observability Elements and Metadata



Data observability serves as the monitoring foundation for DataOps, which is an established discipline for building and managing data pipelines. DataOps applies principles of DevOps, agile software development, and total quality management to data pipelines to help deliver timely, accurate data to the business. DataOps comprises testing, continuous integration and deployment (CI/CD), orchestration. Data observability delivers intelligence that makes each of these DataOps components more effective (see figure 2).

Figure 2. Data Observability as the Monitoring Foundation of DataOps



Challenges

Data observability programs face the challenges of rising data volumes, noise, isolated signals, fragmented tools, and siloed teams.

- **A data explosion.** Data pipelines can choke on rising volumes, varieties, and velocities of data, driven by proliferating sources and digitized interactions.
- **Noise.** Pipelines and infrastructure emit millions of logs and metrics, creating noise that drowns out the signals of data health.
- **Isolated signals.** Data teams, data engineers in particular, struggle to connect isolated signals so they can understand sequences of events.
- **Fragmented tools.** Many tools give fragmented views. For example, application performance management tools focus on operational rather than analytical workloads.
- **Siloed teams.** Business and IT stakeholders struggle to collaborate because they lack common views and ways to communicate.

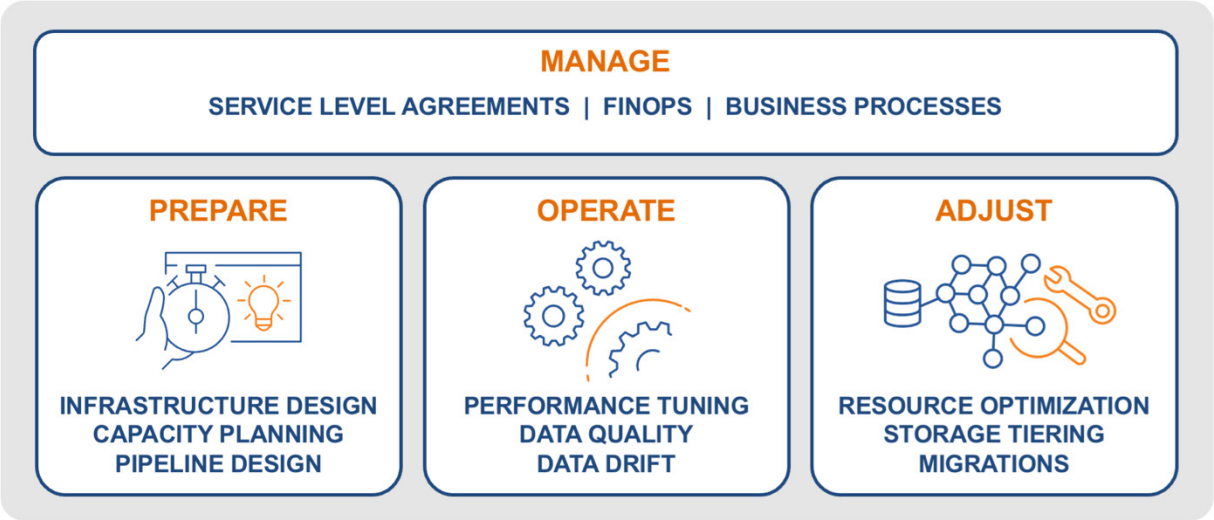
Benefits

Enterprises that overcome these obstacles can build data observability programs that increase the upside of data and reduce its downside. Data teams can increase their productivity, delivering higher volumes of more accurate data, and increase efficiency of the infrastructure resources that support them. Business teams can count on better uptime for their analytics projects and applications, with lower risk of missing service level agreements (SLAs). They also can count on analytics teams to generate higher value and greater agility for the business as they compete in dynamic markets.

How to Observe Data: Use Cases

Now we explore the use cases for data observability, including their requirements and techniques for meeting those requirements. These ten use cases fall into four categories: preparing your data environment, operating it, adjusting it, and finally, funding analytics projects and applications (see figure 3.)

Figure 3. Use Cases for Data Observability



Prepare

Data observability helps data teams prepare the data environment. The primary use cases for this category are designing their infrastructure, planning the capacity of infrastructure resources, and designing the pipelines that deliver data for consumption.

Data Infrastructure Design

Data architects and engineers need to design high-performance, flexible, and resilient data architectures that can meet SLAs. To achieve this, they need to understand performance and utilization trends for their infrastructure. Perhaps slow network connections tend to delay ML models for fraud prevention, increasing customer and merchant wait times for large transactions. Perhaps financial analysts have long wait times when building earning reports because business analysts swamp the CRM database with ad-hoc queries.

Data observability tools measure events and trends like these, helping data architects and engineers design the right infrastructure to support the next round of analytics projects. They might decide to place their merchant records in a cloud data store so that elastic compute can support fraud-check queries at low latency. They might create a secondary CRM database to eliminate resource contention between business analysts and financial analysts, or a separate repository for unused data. These are just some of the ways in which data observability gives data teams the intelligence they need to design data architectures that will support their analytics projects and applications.

Data observability tools help data architects and engineers design infrastructures based on analysis of performance and utilization trends.

Capacity Planning

Once data architects and data engineers select the infrastructure elements for their pipelines, platform engineers must work with CloudOps engineers to plan their capacity needs. They must provision the right resource amounts, maintain healthy utilization levels, and request the corresponding budget. Data observability helps in several ways. For example, engineers can simulate workloads to predict the necessary memory, CPU, or bandwidth required to handle a given workload at a given SLA. This helps them define an appropriate mix of resources, with reasonable buffer capacity for growth, and avoid spending money on unnecessary resources. Data observability also helps measure and predict the variance of key performance indicators (KPIs) to define points at which workloads risk becoming unstable.

Pipeline Design

Data architects and data engineers need to design flexible pipelines that can extract, transform, and load data from source to target with the appropriate latency, throughput, and reliability. This requires a granular understanding of how pipeline jobs will interact with elements such as data stores, containers, servers, clusters, and virtual private clouds. Data observability can help.

Data observability provides a granular understanding of how pipeline jobs will interact with infrastructure elements such as data stores, containers, and clusters.

Specifically, data engineers can profile workloads to identify unneeded data, then configure a pipeline job to filter out that data before transferring it to a target. Profiling workloads also enables them to determine the ideal number of compute nodes for parallel processing. In addition, they can identify hotspots for existing pipelines—i.e., over-utilized platforms or resources—and schedule jobs to avoid such risks when designing the next pipeline. Along similar lines, they might identify table joins or other inefficient processes within existing pipelines and avoid them in future pipelines. All this helps data engineers build effective pipelines and smooth their rollouts into production.

Operate

Data observability also helps data teams operate their environment. The use cases for this category include studying and tuning pipeline performance, finding and fixing data quality issues, and identifying data drift that affects machine learning (ML) models.

Performance Tuning

When a production BI dashboard, data science application, or embedded ML model fails to receive data on time, decisions and operations suffer. To prevent such issues and mitigate their damage, data and CloudOps engineers must tune their pipelines based on health signals such as memory utilization, latency, throughput, traffic patterns, and compute cluster availability. Data observability tools can help them do this, improving their ability to meet SLAs for performance.

For instance, a data engineer might configure threshold-based alerts to flag pipeline errors or delays, such as the failure to deliver an expected dataset as scheduled. When they receive this alert through a notification system such as **Slack**, they pull up a pipeline diagram that shows service level indicators and KPIs for each sequential job. Then they drill into relevant logs, metrics, or traces to identify the root cause. Perhaps a **Spark**-based EC2 cluster was over utilized, which slowed down the transformation of a customer dataset before it arrived in **Delta Lake** for processing by an ML model. Responding to an AI-driven recommendation, the data engineer collaborates with a CloudOps engineer to remediate the issue by increasing the size of the cluster.

Data Quality

Data quality is the foundation of success for analytics. Without accurate views of the business, a sales dashboard, financial report, or ML model might do more harm than good. To minimize this risk, data teams need to find, assess, and fix quality issues with both data in flight and data at rest. The objective is to resolve quality issues before business owners or customers find them. Data observability can help.

Data observability can help resolve data quality issues before business owners or customers find them.

Say that a data observability tool scans and profiles a target table, then recommends ML-driven rules to detect anomalies in similar tables or updates to that table. These anomalies might include missing values, duplicates, or out-of-range values. The data engineer reviews and selects the right rules, possibly adding their own custom rules, then schedules periodic or continuous quality checks based on them. They also create aggregate quality scores for tables or other data assets, based on the results of rule checks, and configure alerts that fire when quality scores break specified thresholds. To provide additional stability, the data engineer monitors any changes in source schemas—for example, when a database administrator changes columns in a table—that might break target applications or algorithms.

Data Drift

ML models tend to degrade over time, meaning that their predictions, classifications, or recommendations become less accurate. Data drift—i.e., changes in data patterns, often due to evolving business factors—can cause much of the degradation. These factors might include the health of the economy, price sensitivity of customers, or competitor actions. When they change, so does the data feeding into ML models, making the models less accurate. Data scientists, ML engineers, and data engineers must identify data drift so they can intervene and adjust the ML models. Data observability can help them do this.

Suppose that an e-commerce company implements an ML model that recommends discounted products for customers to add to their shopping cart. The model bases the discount levels on purchase histories, in particular, how much of a discount was required to drive sales in the past. Then the economy slows, making customers less willing to purchase. A data observability tool helps data science teams spot the problem fast by alerting them to changes in purchase indicators such as sales win rates, transaction values, and total revenue. The ML engineer or data scientist then intervenes to adjust the ML model by re-training it on fresh data and increasing the discounts.

Adjust

Next, data observability helps data teams adjust the data environment. This category includes use cases such as resource optimization, storage tiering, and migrations.

Resource Optimization

Analytics and data teams can make ad hoc changes that lead to inefficient resource consumption. For example, the data science team might decide to quickly ingest huge new external datasets so they can re-train their ML models for customer recommendations. The BI team might start ingesting and transforming multi structured data from external providers to build 360-degree customer views. Data and CloudOps engineers tend to support new workloads like these by consuming cloud compute on demand, which can create cost surprises. Data observability helps optimize resources to get projects back within budget.

Data observability helps optimize resources to get projects back within budget.

A data observability tool can model new workload requirements, identify servers on premises that can handle the workloads, then recommend scheduling them to run there during off hours. By accepting this recommendation, data science and BI teams can train their models and transform their data

without consuming costly compute from a cloud service provider. The data observability tool can help balance workloads in other ways, for example by placing long-running jobs on premises while keeping low-latency applications in the cloud. It also can recommend unnecessary memory usage to trim, and duplicate datasets to consolidate.

Storage Tiering

While analytics projects and applications require high volumes of data, they often focus on just a small fraction of it. Once trained, an ML model for fraud detection might need just 10 columns out of 1,000—the so-called "features"—to assess the risk of a given transaction. Once implemented, a sales performance dashboard might need just a few fresh data points each week to stay current. In both these cases, the rest of the data remains "cold," with few if any queries. Data observability can help data engineers identify cold data and offload it to a less expensive tier of storage.

To illustrate, a data observability tool can help inspect and visualize "skew," which refers to the distribution of I/O across columns, tables, or other objects within a data store. A data engineer might find that just 10% of the columns or tables in a CRM database support 95% of all queries. They might also find that most sales records aged more than five quarters are never touched again. Based on these findings, the data engineer creates automated rules to archive this cold data in a cheaper repository such as **Amazon Glacier**.

Migrations

Cloud migration initiations can create the need for many of the use cases described here. In addition, data observability can assist migrations themselves. Before migrating analytics projects to cloud platforms, data and CloudOps engineers need to answer some basic questions. What does the target cloud environment look like? How well will that environment support their analytics tools, applications, and datasets? And how will their analytics workloads perform in that environment? Engineers that fail to answer questions like these in advance might derail their migration or undermine analytics results afterwards. Data observability can help answer such questions and minimize the risks.

For instance, a data observability tool can help discover clusters and other infrastructure elements to create an inventory of assets in their target cloud environment. Data and CloudOps engineers can use this inventory to compare source and target environments in detail, then identify risks such as incompatible tools, APIs, applications, data formats, or security settings. They also model how existing workloads would behave in the target environment, and what that would mean in terms of resource requirements and cost. During the migration itself, data observability tools can monitor various KPIs to help maintain application uptime and minimize the risk of business disruption.

Fund

Finally, data observability helps business and IT leaders fund analytics projects and applications from a business perspective. This category of use cases focuses on the FinOps use case.

Financial Operations (FinOps)

Cloud platforms give enterprise financial flexibility, enabling them to rent elastic resources on demand rather than buying servers and storage arrays for their own data centers. But those elastic resources, compute in particular, can lead to budget-breaking bills at the end of the month. Surprises like these have given rise to FinOps. This emerging discipline helps IT and data engineers, finance managers, data consumers, and business owners collaborate to reduce cost and increase the value of cloud-related projects. FinOps instills best practices, automates processes, and makes stakeholders accountable for the cost of their activities. Data teams use FinOps to make cloud-based analytics projects and applications more profitable, and they need the intelligence of data observability to achieve that.

The emerging discipline of FinOps needs the intelligence of data observability to manage the cost of cloud-based analytics.

To illustrate, a data or CloudOps engineer can use a data observability tool to track operational costs based on compute, storage, and network usage. They can measure consumption by application, project, or team to charge costs back to accountable parties. They also can model workloads to forecast the cost of future workloads and SLAs, helping IT plan their budgets and business owners plan their analytics projects. In these and other ways, data observability helps guide the decisions and actions of various FinOps stakeholders.

Guiding Principles

Planned and implemented well, a data observability program helps enterprises build new trust in their analytics projects and applications. By monitoring both data pipeline performance and data quality, data teams can optimize the timeliness and accuracy of data delivery. They can address four categories of use cases: preparing their data environment, operating it, and adjusting it, as well as managing the business aspects of their analytics projects and applications. Eckerson Group recommends that data leaders use the following guiding principles to increase the odds of success with a data observability program.

- **Let the business prioritize your use cases.** Each use case described in this report costs time, effort, and money. Engage business owners as you assemble your use cases and identify which ones to address. These business owners, and the data analysts and data scientists that work for them, can help prioritize where to start based on a cost-benefit analysis of each use case. Aim for demonstrable results with low risk.
- **Foster team collaboration.** These use cases require multiple IT and business stakeholders to collaborate. Data engineers, platform engineers, CloudOps engineers, data analysts, data scientists, and finance managers must learn to communicate and help one another. This might require training, dotted-line reporting structures, and periodic cross-functional meetings to ensure stakeholders pass the right batons at the right times.
- **Standardize on shared tools.** Teams collaborate more effectively when they share common tools. Look for a multi-modal platform for data observability that both IT engineers and analytics teams can use—and that business owners can oversee. The more your stakeholders can reduce friction by changing between tools and interfaces, the faster they can solve problems and deliver business value.

About Eckerson Group

Wayne Eckerson, a globally-known author, speaker, and consultant, formed **Eckerson Group** to help organizations get more value from data and analytics. His goal is to provide organizations with expert guidance during every step of their data and analytics journey.

Eckerson Group helps organizations in three ways:

- **Our thought leaders** publish practical, compelling content that keeps data analytics leaders abreast of the latest trends, techniques, and tools in the field.
- **Our consultants** listen carefully, think deeply, and craft tailored solutions that translate business requirements into compelling strategies and solutions.
- **Our advisors** provide one-on-one coaching and mentoring to data leaders and help software vendors develop go-to-market strategies.

Eckerson Group is a global research and consulting firm that focuses solely on data and analytics. Our experts specialize in data governance, self-service analytics, data architecture, data science, data management, and business intelligence.

Our clients say we are hard-working, insightful, and humble. It all stems from our love of data and our desire to help organizations turn insights into action. We are a family of continuous learners, interpreting the world of data and analytics for you.

Get more value from your data. Put an expert on your side. **Learn what Eckerson Group can do for you!**

Eckerson Group

RESEARCH CONSULTING ADVISORY

GET • MORE • VALUE • FROM • YOUR • DATA

About the Sponsor

Acceldata is the market leader in enterprise data observability for the modern data stack, Founded in 2018, Campbell, CA-based Acceldata has developed the world's first enterprise Data Observability Cloud to help enterprises build and operate great data products. Acceldata's solutions have been embraced by global customers, such as Oracle, PubMatic, PhonePe (Walmart), Verisk, Dun & Bradstreet, DBS, and many more. Acceldata investors include Insight Partners, March Capital, Lightspeed, Sorenson Ventures, and Emergent Ventures.

We invite you to visit us at **Acceldata.io** and follow us on **LinkedIn** and **Twitter**.