Bigeye

State of data quality report

Introduction

Just as data needs monitoring, so too does our understanding of how organizations handle data quality. To provide “observability for data observability”, we conducted a qualitative and quantitative survey of data practitioners. We combined survey results with other qualitative and quantitative research recently conducted on the subject of data quality monitoring.

Of 100 survey respondents, at least 63 came from mid-to-large cloud data warehouse customers (with a spend of more than $500k per annum) who have some form of data monitoring in place, whether third-party or built in-house.

The result, Bigeye’s 2023 State of Data Quality report, sheds light on the perennial scourge of data quality, and how data practitioners believe it must be addressed.

Survey results reveal that data quality and reliability continue to pose significant challenges for organizations, impacting customers and overall productivity. Despite the efforts of data engineers, software engineers, and data analysts, who are typically responsible for data issues, issues still take anywhere from 1-2 days to weeks and even months to spot and fix. More than half of the respondents have experienced five+ data issues in the last three months.

Alarmingly, 20% of respondents have faced at least two severe data incidents in the past six months that have directly impacted the business's bottom line and drawn attention at the highest levels of the organization.

These findings underscore the need for automated solutions like data observability and tooling, as well as organizational and process improvements that break down communication barriers between consumers and producers of data.

This report serves as a call to action for organizations to recognize the critical importance of data quality, and we hope it provides valuable insights for data leaders to make informed decisions about their data strategies.

Section 1

Executive summary/ findings

86%

Of respondents reported that some combination of data engineers, software engineers, and data analysts are in charge of data at their organization.

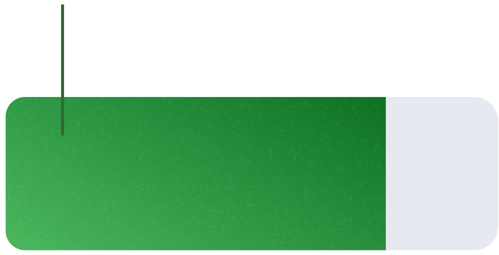
52%

Of respondents experienced more than 5 data issues in the last three months.

40%

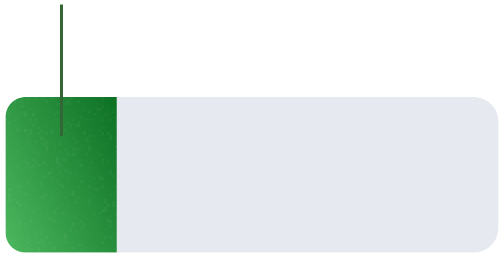
Of respondents reported at least two “moderately severe” data incidents in the last six months, which were prevented from creating damage to the business/bottom line only with heroic effort.

78%



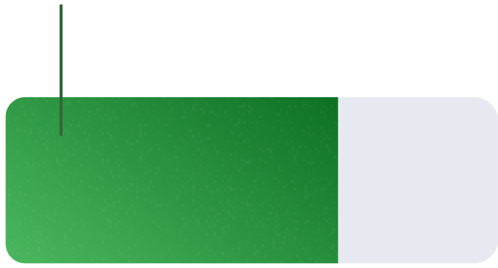
Of respondents reported that some combination of data engineers, software engineers, and data analysts are responsible for fixing data incidents when they arise. BI analysts/product teams also assist.

20%



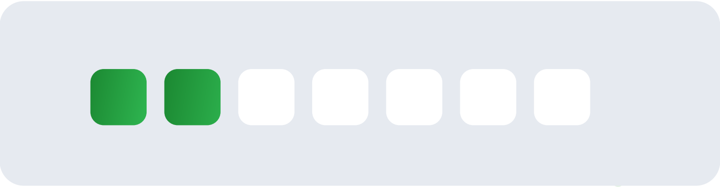
Of respondents reported at least two “severe” data incidents in the last six months, which created damage to the business/bottom line and were visible at the C-level.

70%



Of respondents reported at least two data incidents that diminished the productivity of their teams.

~1-2 days



Data issues most commonly take ~1-2 days to spot and fix, but with a long tail lasting up to weeks and months.

Section 2

Rise of the data engineer

Rise of the data engineer

Research pinpointed the rise of the “data engineering” role, which is now as popular as the "data scientist" role.

After a decade of “data science” repeatedly topping “hottest jobs” lists, those roles are now being joined by others. They are data engineers (in charge of managing data pipelines and data quality) and data analysts/business analysts (consuming the data, either by building dashboards or by using the data to drive business decisions).

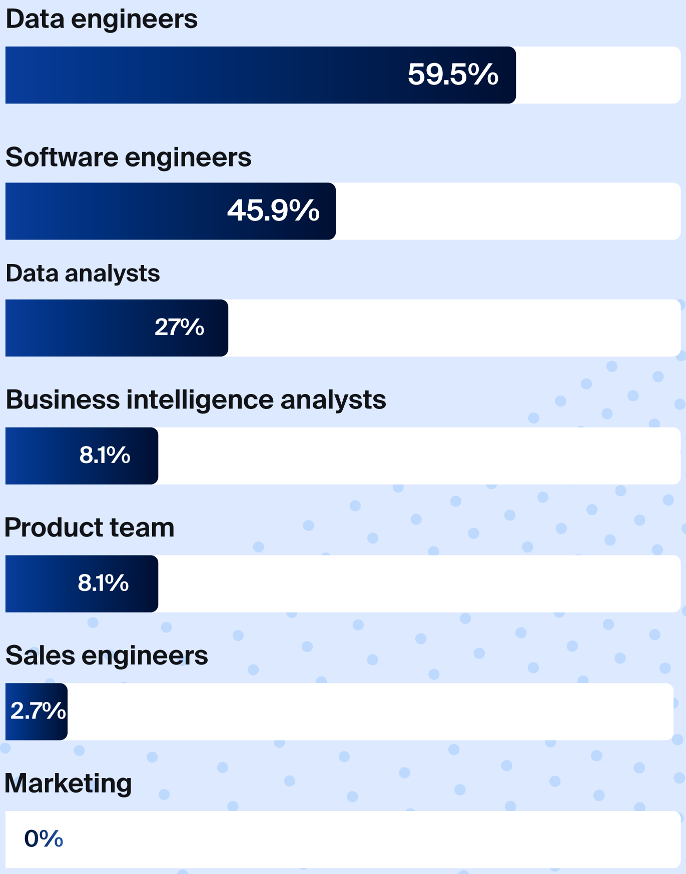
Our survey found that data engineers are the first line of defense in managing data issues, followed closely behind by software engineers.

The role of data engineer has now moved closer to that of software engineer. Like software engineers, data engineers are in charge of a product - the data product - that increasingly demands software-like levels of process, maintenance, and code review.

New disciplines like data reliability engineering aim to bring best practices from traditional software engineering (think observability and site reliability engineering) to bear on the data product.

Data quality work is now largely the responsibility of data engineers and software engineers, with smaller contributions from data analysts.

When data problems ocurr, who is the first line of defense in managing them?



Section 3

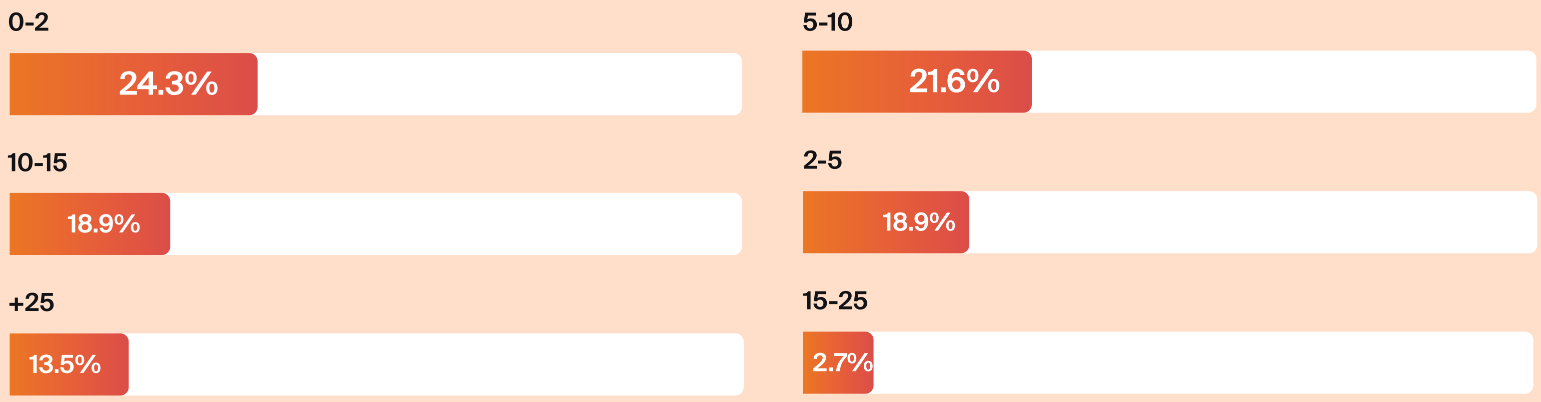
Data quality incident frequency

The frequency of data quality incidents

Research revealed that companies are experiencing a median of **5-10 data quality incidents over a period of three months.**

These incidents range from severe enough to impact the company's bottom line, to (merely!) reducing engineer productivity. In other words, that's 500 hours of data downtime per quarter that's impacting teams.

How many data incidents have you experienced in the past three months?



Moreover, **15% of respondents report more than 15 data incidents in the past three months.**

Lastly, our research found that **data quality issues take an average of 48 hours** to fully troubleshoot.

Organizations with more than five data incidents a month are essentially lurching from incident to incident, with little ability to trust data or invest in larger data infrastructure projects. They are largely performing reactive over proactive data quality work.

For instance, an executive looks at a dashboard, notices that the revenue number is too low, and asks why. An individual data engineer or software engineer then spends hours debugging the data pipeline to figure out why. The next week, the cycle repeats.

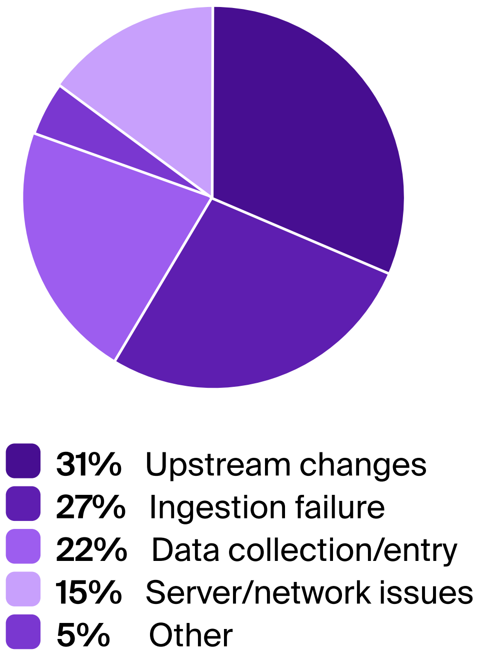
Section 4

Data quality solutions require both technology and process improvements

Data quality solutions require both technology and process improvements

According to our survey, data quality issues fall into a few categories:

Some of these challenges require technological solutions; others cannot be fixed except by changes to organizational processes.



> “Downstream data eng work not done (correctly) when the source data model changes, unclear definitions/docs, legacy systems are not migrated leading to mistakes.”

> “Most severe is probably some application-side change to a model that isn't accounted for in the warehouse schemas”

> “Application code changes”

Upstream changes

At a growing company, it’s impossible to expect schemas, data types and formats, and applications to stay static: there are always upstream changes. Issues happen whenever changes aren’t properly communicated to the downstream data consumers.

To improve communication between consumers and producers of data, organizations can take a number of actions.

The fix: Automations

As a lightweight solution, companies might implement Github automations that tag PRs involving data model changes with reviewers from the consuming team.

The fix: SLAs

On the more comprehensive side, data SLAs and data contracts specify formal commitments to the data’s framework and quality. There are penalties for violations.

Data collection/entry

Data collection/data entry errors plagued a significant amount of survey respondents.

A typical data entry error is as follows: your application has a form page, from which it collects user email addresses. However, the form doesn't have proper validation and checks. Users end up accidentally (or not so accidentally) typing in emails in the wrong format, or with incorrect information.

The fix: Robust ELT pipelines

Data collection errors can be remedied by ELT; the more robust the better for data engineering teams. They might write the ELT pipelines, then work with the product engineering team, then implement form validations on the application frontend.

"Bad collection (e.g., missing records, messy records, etc.)"

"No internal standards for application output"

"Insufficient validations and schemas"

"Customer filled in the data incorrectly"

"Data entry issues not handled by ETL"

Ingestion failure

Data pipeline/infrastructure issues accounted for a significant percentage of respondents' data issue root causes. These issues typically involve source data not being where they should be at the prescribed time and are caused by certain parts of the pipeline failing, with the effects cascading down the pipeline.

Ingestion failure happens commonly because data isn't stationary. As it flows through pipelines in scheduled workflows, errors occur. Teams are finding that it's easier said than done to receive "good quality" data at the end of the pipeline. It requires that each segment run correctly AND operate on correct inputs.

"Ingest running without all the pre-requisite data being ready"

Server/Network issues

A variety of server/network issues plague data quality across respondents. Server and network issues can lead to duplicate data, unstructured data, incomplete data, different data formats, and/or difficulty accessing the data. One respondent noted that infrastructure and network problems lead to “unanticipated changes in production that break data replication”.

The fix: Data governance

In the context of an organization, “governmental regulation” amounts to universally agreed-upon expectations and rules around data, with the rights and responsibilities of each party outlined clearly. Collectively, this is referred to as “data governance”.

Software engineers and data engineers feel disempowered

Survey results highlighted that engineers and data engineers often feel disempowered when dealing with data quality issues.
A variety of factors are at fault:

- Lack of incentive
- Lack of visibility into the root cause
- Lack of ownership

“Refresh timeouts”

“API failure”

“Issues with infrastructure, like via Tableau”

“Third-party infrastructure”

“Everyone needs a scorecard because no engineer will build this unless they have to”

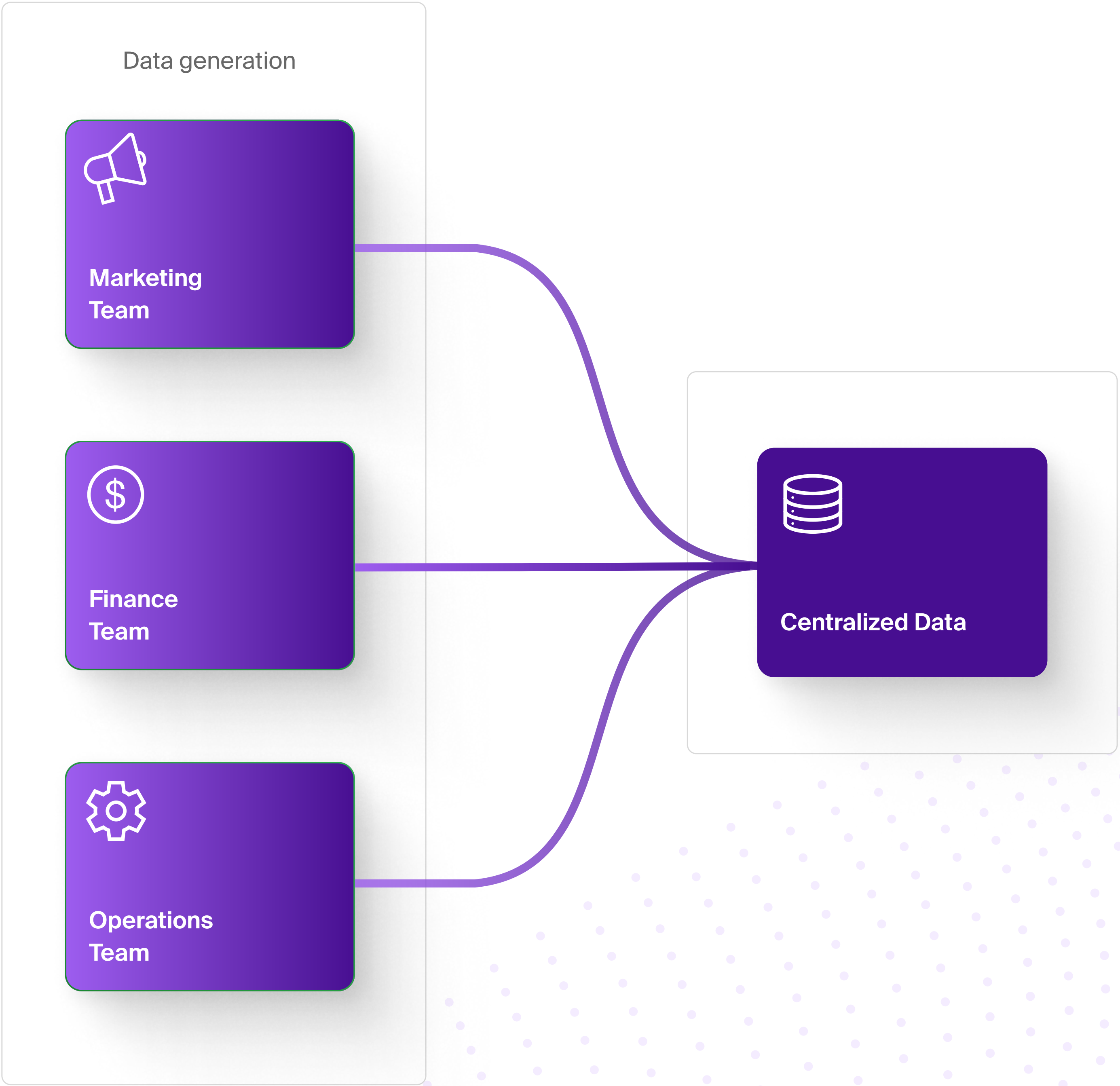
“We have an overwhelming number of datasets with lack of clarity as to what is important/who should do the work”

“The problems are mostly the same, but not a space in which I have control or visibility”

DATA QUALITY SOLUTIONS REQUIRE BOTH TECHNOLOGY AND PROCESS IMPROVEMENTS

Historically, data quality initiatives are difficult to execute since data production and consumption are spread out and shared across an entire organization.

In other words, data quality is the ultimate tragedy of the commons: when each data user or producer simply acts in their own self-interest, they're incentivized towards actions like duplicating tables and producing untidy data, actions that complicate and deteriorate the data product.



Section 5

Desire for automation

Desire for automation

Research found that engineers are looking for proactive, automated solutions to address data quality issues. In most companies, data quality initiatives are generally prioritized after a painful data quality incident that negatively impacts the bottom line. This means that solutions are too often reactive point fixes.

In general, respondents are looking for more systematic approaches to data quality that don't rely on institutional memory of data models or application logic.

In-house solutions lack automation and are worse at error detection

Research showed that in-house solutions can:

- Require larger maintenance overhead
- Be too expensive on warehouse compute
- Be ill-suited for balancing monitoring needs and warehouse costs
- Be worse at identifying data inconsistencies

Respondents who used third-party data monitoring solutions found approximately a 2x to 3x ROI over in-house solutions. They also noted that at full utilization, third-party data monitoring solved for two issues: fractured infrastructure, and anomalous data. They further reported that third-party data monitoring solutions had better test libraries, and a broader perspective on data problems.

"Make it programmatic. Remove humans"

"It's difficult to maintain complex joins of multiple records"

"For us, it's much less about the technical elements of data quality (anomalous values, broken pipelines) and much more about our logic being complicated, and the origin and nature of some issues being hard to pinpoint"

Respondents told us that it would take an average of 37,500 man-hours to build an in-house data quality monitoring solution. Roughly, that equates to one year of work for approximately 20 engineers.

Data quality monitoring is already here, and third-party solutions win

89% of respondents indicated that data monitoring is either somewhat critical or very critical to their operations.

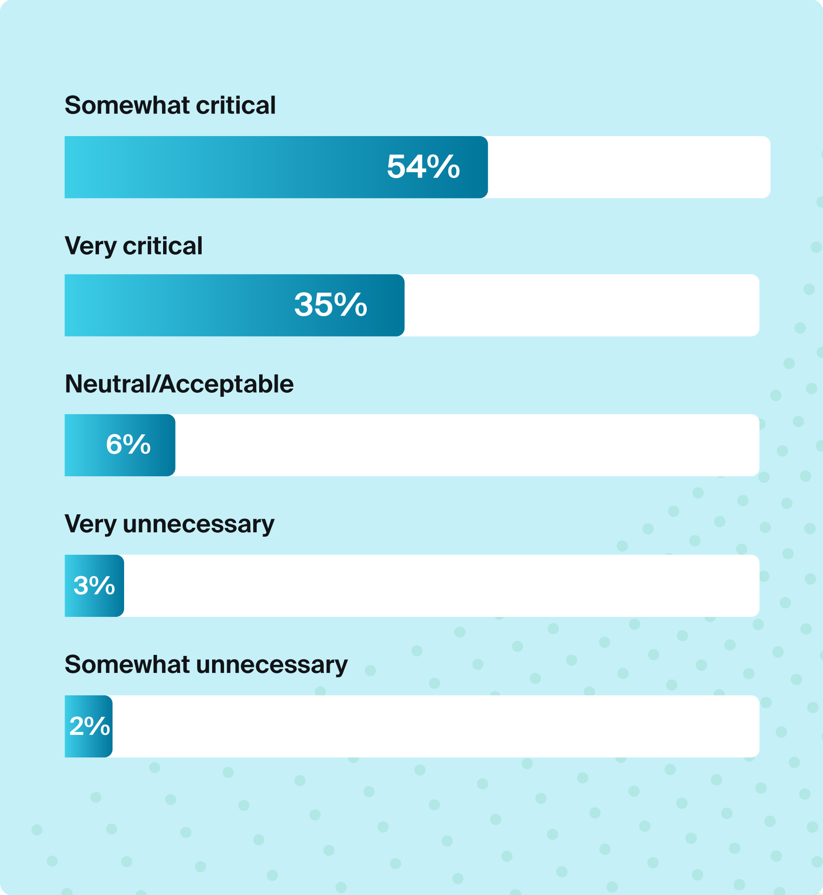
Of 25 respondents who had used both an in-house and a third-party solution, 100% said that the third-party solution reduced monitoring man-hours.

40% said that the third-party solution saved 30% or more of their time.

Third-party solutions were reported to have intangible benefits like reassurance and security. While not quantifiable ROI, those benefits often prove to be just as valuable.

Several respondents noted the value of relying on third-party data quality monitoring platforms for credibility in escalating data issues. They also noted that having automated data quality monitoring running in the background provides org-wide reassurance and peace of mind.

56% of respondents indicated that reducing compliance and regulatory risk was a critical factor in their decision to implement a third-party solution.



Conclusion

There is a huge variety in the solutions that companies leverage to combat data quality issues. They range from data analysts hand-checking data to analysts hacking together their monitoring with scheduled queries to full-blown third-party data observability tools.

This research shows that automation, schema validation, source checks, and comprehensive monitoring are necessary. Gone are the days of haphazardly addressing data quality. Data quality monitoring is here to stay. Going forward, we predict that formal data quality monitoring will grow more comprehensive and become standard as best-practice across most industries that have a technology component.

Want to continue your data quality journey?

Request a demo

Talk to someone from our team to take Bigeye's data quality monitoring tool for a spin.

Check out the Bigeye blog

Get expert insights, self-assessments, interviews, and long-form guides on the latest and greatest in data.

Bigeye

www.bigeye.com