

WAKEFIELD SURVEY

**2023**

---

# THE STATE OF DATA QUALITY

DATA LEADER  
STRATEGIES &  
BENCHMARKS

 MONTE CARLO

# Table of Contents

## **I. Introduction**

- Executive summary
- Methodology

## **II. Modern Data Stack Complexity**

- Number of tables
- Number of dbt models
- Number of data tests

## **III. The Rise of Data Downtime**

- Time to detection
- Time to resolution
- Number of incidents

## **IV. The Cost of Data Downtime**

- Time spent on data quality
- Business stakeholder impact
- Revenue impact

## **V. Ownership and Solutions**

- Primarily responsible for data quality
- Hours to build ML data monitoring tool
- Most likely reason to launch an initiative

## **VI. Additional Resources**

- Resources

# Executive Summary

Before launching the industry's first end-to-end data observability platform, we spoke with hundreds of data teams about their biggest pain point: data downtime.

The problem of erroneous and inaccessible data still exists today. In fact, [data downtime nearly doubled year over year](#).

This was driven by a [166% increase](#) in the average time to resolution.

The consequences of data downtime are severe, more so than ever before. And in this economic climate, organizations can't afford the cost.

We saw them in the headlines when Unity Technologies, for example, cited a \$110M decrease in their yearly revenue target as a result of bad customer data during their Q1 earnings call.

We saw them in this survey too. Respondents reported [poor data quality impacted 31% of revenue](#) and [74% said data consumers identified issues first](#) "all or most of the time." Both figures increased year over year.

These consequences weren't felt evenly, however.

For instance, organizations with 100 tables or fewer don't spend as much time on data quality as their larger peers. Their data downtime, and associated consequences, are more severe as a result.

Organizations with more tables have improved their operational response. However, they are spending a considerable amount of their team's resources on data quality to do so.

Data leaders shouldn't have to choose between spending half of their team's resources on data quality or facing dire consequences from data incidents.

We believe data observability is the answer. Automatic monitoring end-to-end across the data stack with machine learning paired with data lineage and other tools to accelerate data quality incident resolution.

We hope you find the survey results as interesting as we did, and that these findings help inspire you to build more reliable, scalable data systems. Until then, here's wishing you no data downtime!

**Barr Moses**  
**CEO & co-founder**  
**Monte Carlo**



# Methodology



The Monte Carlo Survey was conducted by Wakefield Research ([www.wakefieldresearch.com](http://www.wakefieldresearch.com)) among 200 data engineers working in the US, between March 9th and March 23rd, 2023, using an email invitation and an online survey.

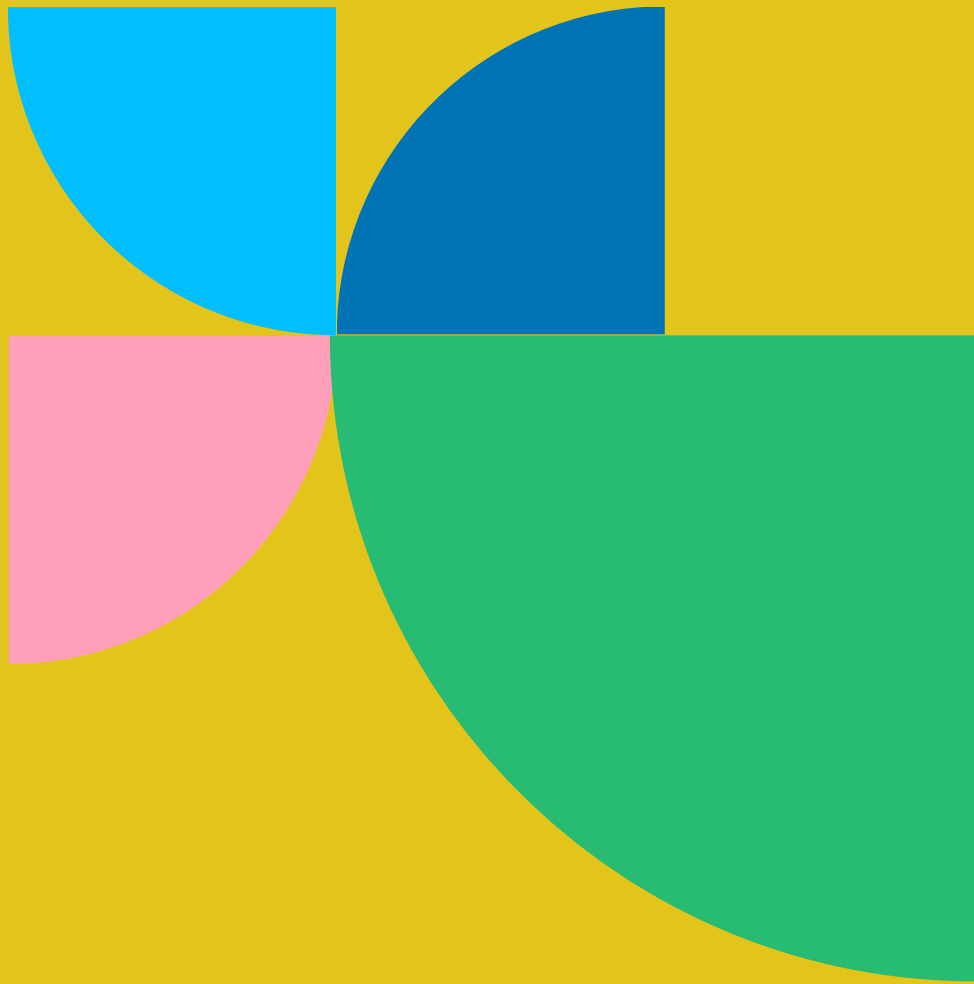
Results of any sample are subject to sampling variation. The magnitude of the variation is measurable and is affected by the number of interviews and the level of the percentages expressing the results.

For the interviews conducted in this particular study, the chances are 95 in 100 that a survey result does not vary, plus or minus, by more than 6.9 percentage points from the result that would be obtained if interviews had been conducted with all persons in the universe represented by the sample.

± Data under “QuickFacts” were derived from the responses, not included as response options that were read during fielding. We include QuickFacts in instances where we feel they will be helpful.



# Section 1. Data Stack Complexity



# Exponential Complexity

The complexity of your data stack expands exponentially by the number of tables.

This is because you aren't just adding a table, but multiple new relationships between that table and others.

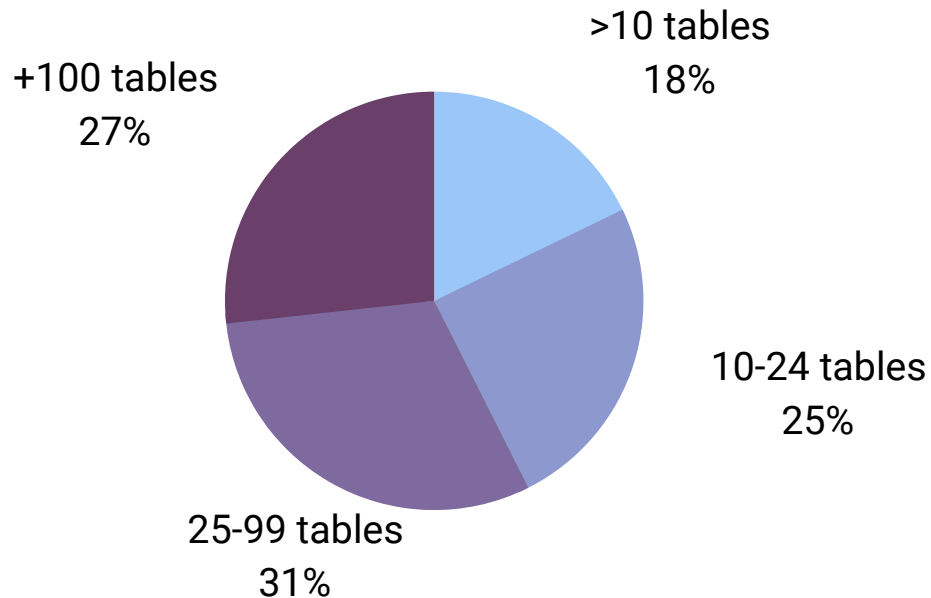
This interdependence is why automatically monitoring across all production tables is ideal. It expedites both detection and resolution.

However, if teams are utilizing a data testing approach they would do best to focus on their most important tables. It doesn't scale otherwise.

## Number of Tables



*How many tables do you have across your data warehouse, lake, or lakehouse?*



*Quick  
Facts*

**AVERAGE TABLES- 642**



**Ryan Kearns**  
**Data Scientist**  
**Monte Carlo**

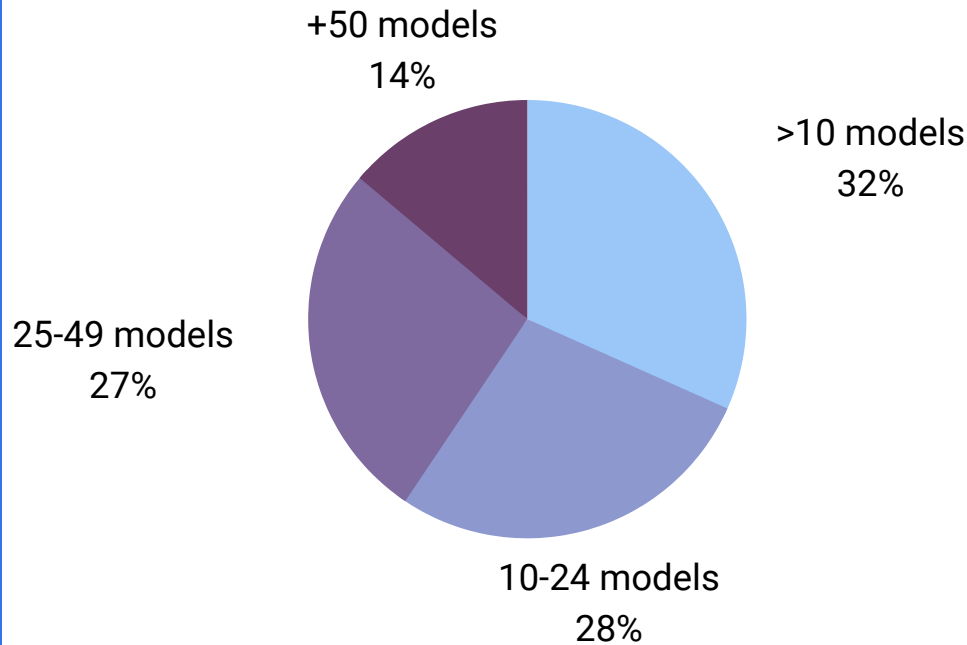
"In modern cloud data infrastructure, few tables stand alone and they rarely contain the root causes of their own data downtime.

When data goes wrong, something upstream in the pipeline – software, data, or schedule – often is wrong, too. This includes software, data, and schedule dependencies, in other words, lineage."

# Number of dbt Models



*How many dbt models or blocks of data transformation code do you have?*



## A Modern Necessity

dbt, or data built tool, is foundational and pervasive in modern data infrastructures.

It accelerates data teams' ability to shape and reshape data to use cases as they arise.

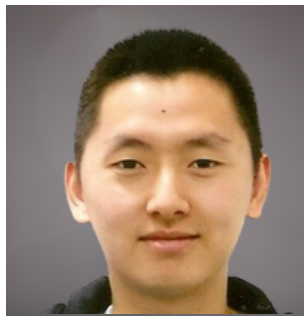
This speed and flexibility can come at a cost however. The more sophisticated and layered your models, the greater the chances for error.

Monitoring models to ensure they run successfully and for how changes to them correlate with incidents is an absolute necessity to minimize data downtime.



*Quick Facts*

AVERAGE MODELS- 24



**Albert Pan**  
**Former Software Engineer**  
**Blend**

“We had many cases where folks would say, ‘Hey, we’re not seeing this data. We’re not seeing these rows. Where are they?’

“It’s never great when a customer tells you that something is wrong or missing. So, we wanted a proactive solution that could tell us when something is wrong, and we can fix it before they even know. That’s one of the main reasons why we use dbt and Monte Carlo.” [[Full Story](#)]

# How Many Is Too Many?

Data testing--either hardcoded or using a tool like dbt or Great Expectations--is a top tactic used by data teams for data quality.

It's also radically insufficient. Humans can't anticipate, and write a test, for all the ways data can break.

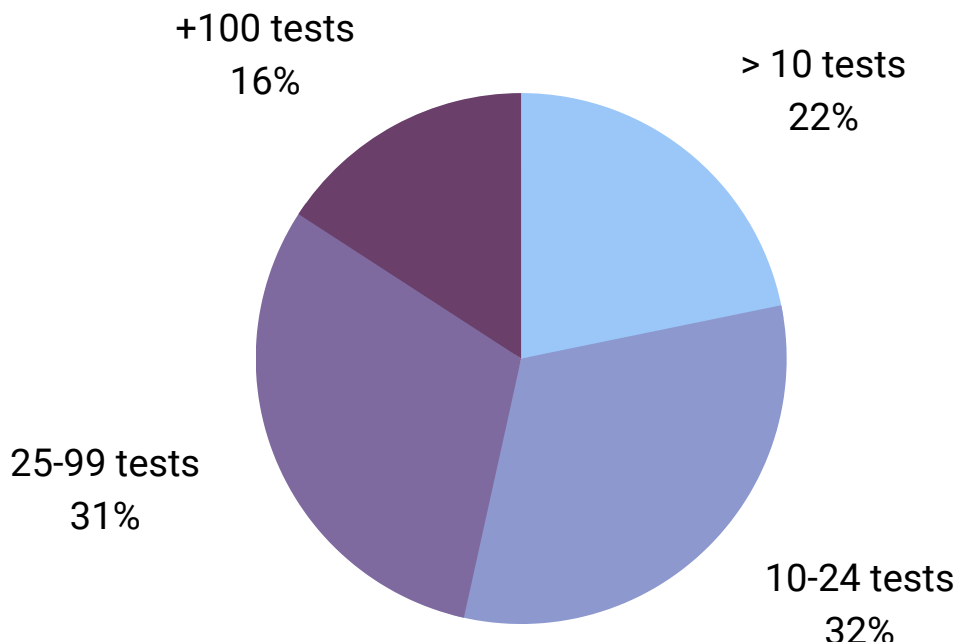
Even if they could, it would be nearly impossible to deploy and maintain these tests at the scale required.

Testing or custom monitors are best deployed on the most important data sets for the most absolute, clear thresholds (ie no NULLs).

## Number of Data Tests



*How many manually written data tests would you estimate you have across all of your data pipelines?*



### Quick Facts

AVERAGE TESTS- 290



**Martynas Matimaitis**  
**Senior Data Engineer**  
**Checkout.com**

“ML-based anomaly detection beats manual threshold basically any day of the week.

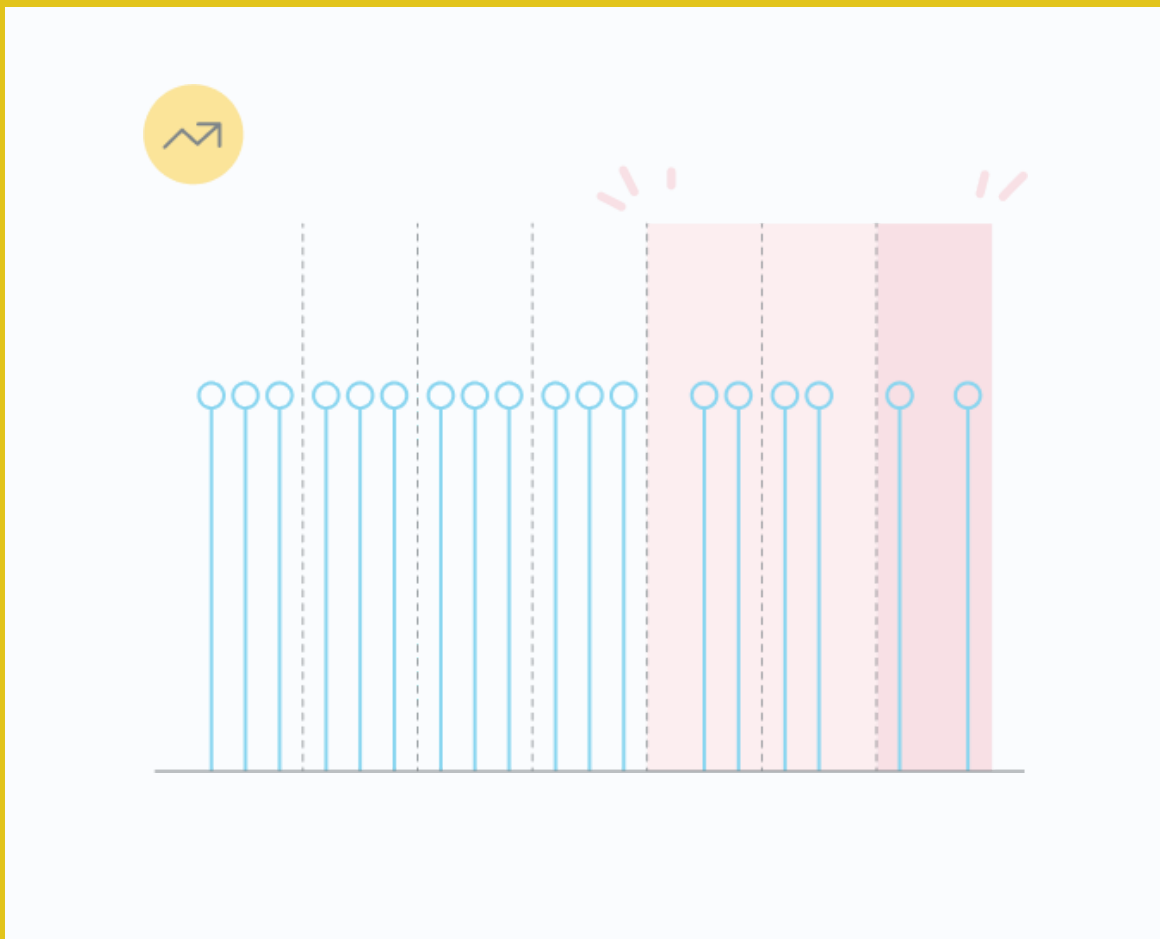
If you actually try to look at [schema changes or anomalies] manually in your entire data warehouse, that’s what takes so much effort to actually capture.

Now, since these models are actually constantly learning and they’re adapting to all the changes and load patterns, over time you get only very few false positives.” [[Full Story](#)].



# Section 2.

## The Rise of Data Downtime



# Time to Detection

## Quick Detection Reduces Risk

The equation for data downtime is [#incidents x (detection + resolution)].

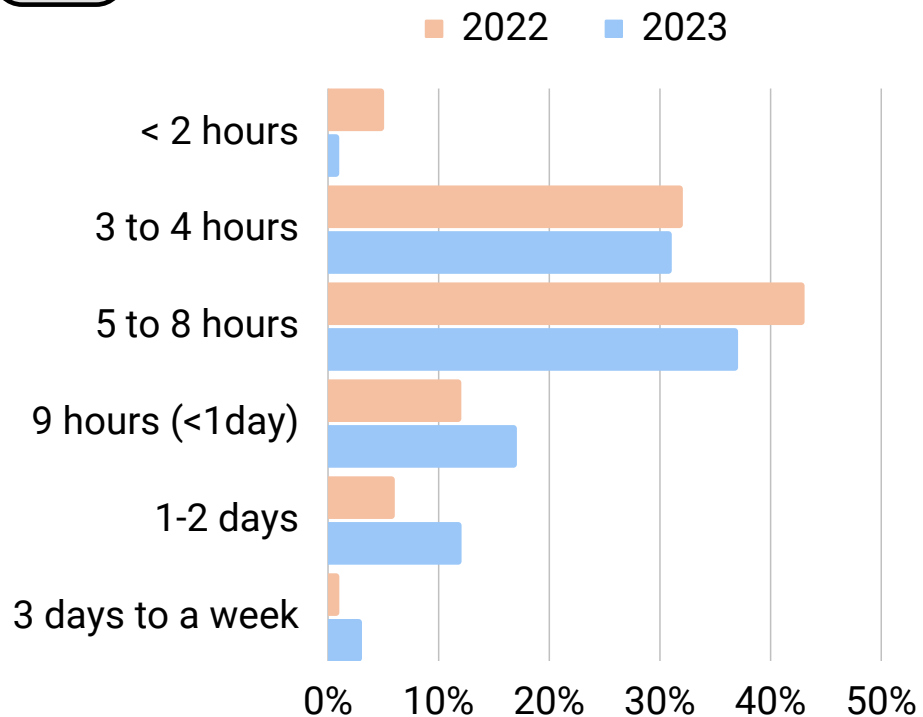
This survey shows all of these variables are directly correlated-- each of these got worse on average for data teams year over year.

In our experience the longer it takes to detect a data incident the more likely it was discovered by someone outside the data team.

Who discovers an incident matters. If it's an internal data consumer it erodes data trust. If it's an external customer that creates a churn risk. And if it's Wall Street...



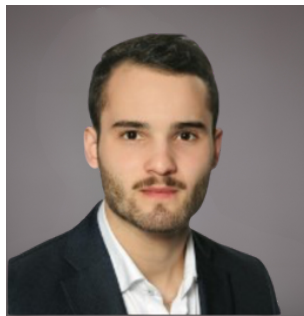
*How long does it typically take to detect a data incident?*



**Quick Facts**

**MORE THAN 4 HOURS-**

**62% to 68% year over year.**



**Otávio Bastos**  
**Formerly Data Governance**  
**Contentsquare**

"We have some unsupervised monitoring, and can

automatically start detecting some very important issues that cannot be detected by human beings....

We really reduced drastically this time to the first response...

[The team is] starting to align themselves within every department to tackle these issues and create a better data environment..." [[Full Story](#)]

# A Looming Crisis

The scale and complexity of data systems have completely overwhelmed data teams.

If teams can't resolve issues efficiently they fall further behind as new ones arise.

Poor time to resolution strongly correlated to more incidents and more time spent on data quality.

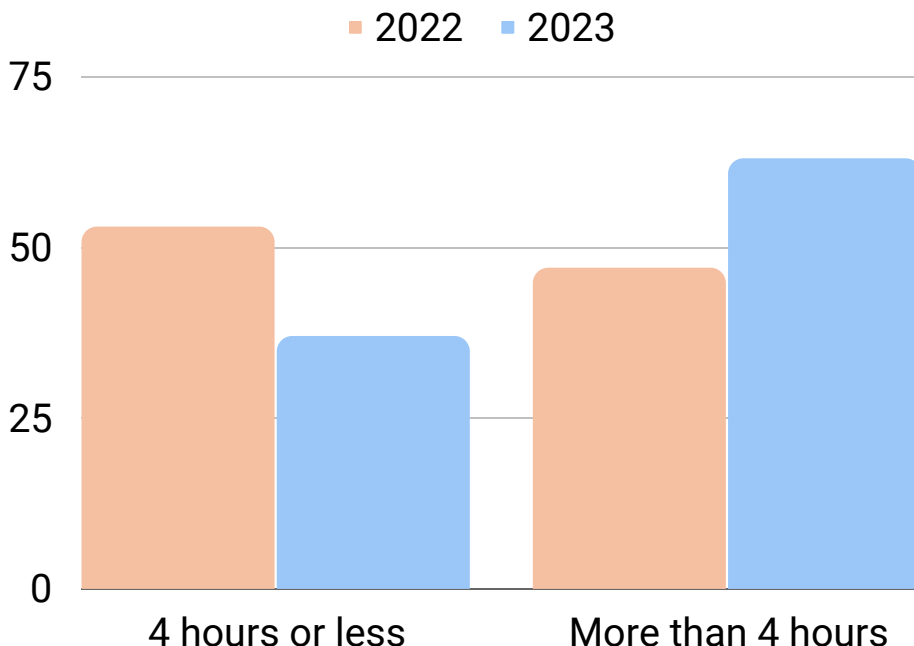
This suggests teams that don't effectively fix incidents continue to suffer issues from the original root cause.

Teams need to prioritize reducing resolution time ASAP.

## Time to Resolution



*Once a data incident is discovered, how many hours does it typically take to resolve the issue or problem?*



**Quick Facts**

Average increased from 9 to 15 hours year over year!



**Kineret Kimhi**  
**Data Engineering Manager**  
**BlaBlaCar**

“When we looked into it we realized that much of our

capacity issues came from data quality issues popping up.

We discovered the process of finding the root cause of data quality issues took 200 hours a quarter across our entire team...

We could see there were failures and we had KPIs, but they couldn't give us the full landscape. That's when we realized we needed data observability ..." [[Full Story](#)].

# What's An Incident Really?

We refer to an incident as a period of time where the data is inaccurate or inaccessible, but teams can get more granular.

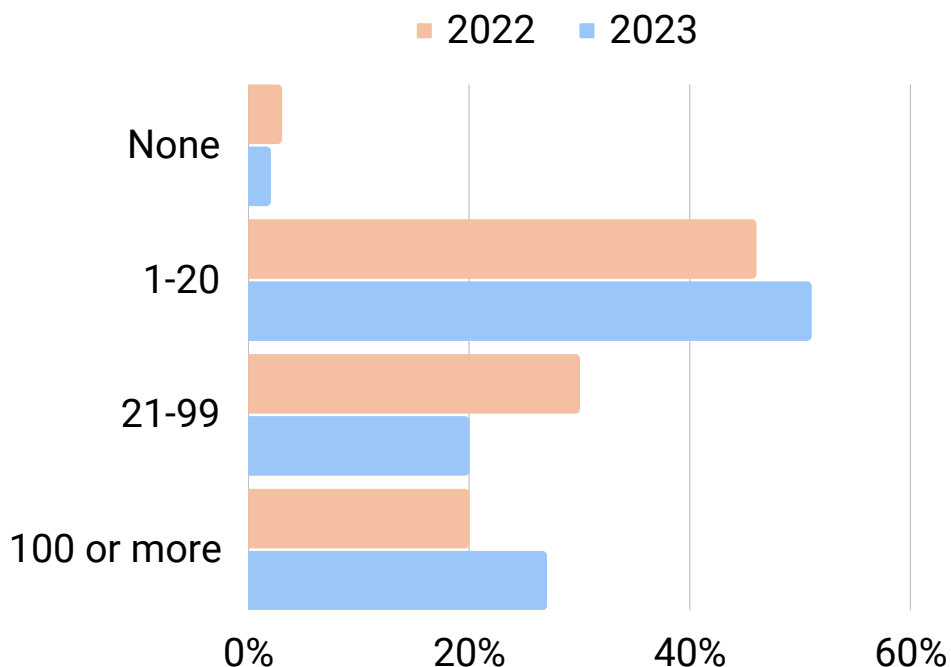
After all not all incidents are created equal. For instance, if a data freshness issue is resolved within an hour before any dashboard user notices--is that an incident?

Data SLAs can give consumers an understanding of what they can expect. They also help data teams place issues in context and better understand where to invest to reduce incidents in the long-term.

## Number of Incidents

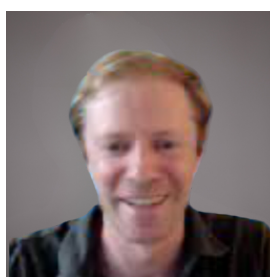


*In a typical month, how many data incidents arise?*



### Quick Facts

Average increased from 59 to 67 incidents year over year.



**Brian London**  
**Director of Data Engineering**  
**SeatGeek**

"One of the things that Monte Carlo has done is enable us to

stabilize our platform. So, in addition to identifying when there is a problem, it has also helped us to understand where problems are likely to occur, where things are brittle.

And over time, we've invested effort into cleaning up our lineages, simplifying our logic." [[Full Story](#)]

# Section 2.

## The Cost of Data Downtime



## Good News?

We want to be happy for teams spending less time on data quality-- it's a benefit of our platform.

The less time engineers are spending fixing, the more time they are innovating and building.

But the question to ask is, "why are data teams spending less time on data quality?"

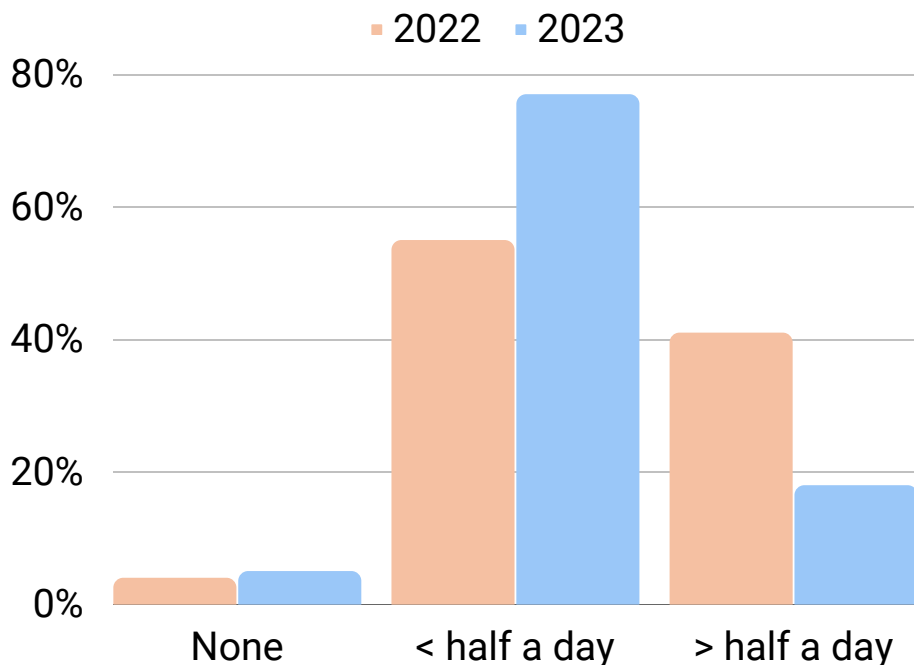
It's only a positive outcome if data quality remains high, otherwise the cost of an incident can greatly outweigh the time saved from sticking your head in the sand.

So which is it? Well, the next few pages offer some clues.

## Time Spent on DQ



*In a typical day, what percent of your time do you spend evaluating or checking data quality?*



**Quick Facts**

Average time decreased from 40% to 34% year over year.



**Brandon Beidel**  
**Director of Engineering**  
**Red Ventures**

"For some teams, more than 50% of all requests were some variant. of a data quality or investigation issue.

Those requests...could trigger a 2 to 3 hour expedition for an engineer... And, unfortunately, the engineers that were the best at finding these issues then became inundated...

We needed to find a way out of this hedonic treadmill and endless loop of time being taken away from productive people." [[Full Story](#)]

## A Business Case No Brainer

Each year data becomes more valuable to organizations. As a result, poor data quality becomes more costly.

It's no surprise then that respondents that reported higher impacts also reported higher levels of data adoption.

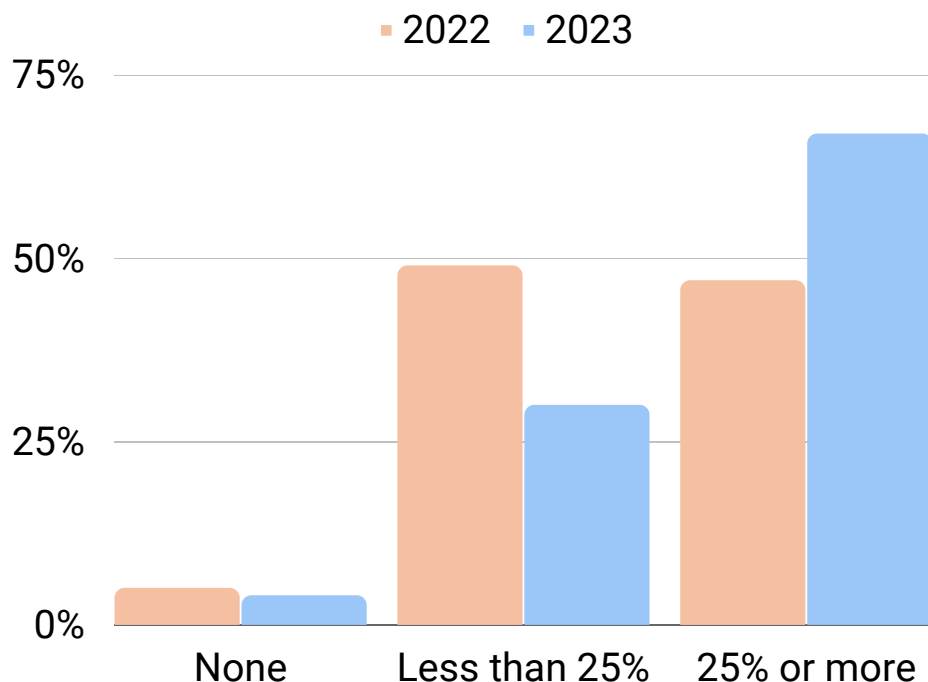
Unfortunately, there was also a correlation with flat budgets as well.

If there is a silver lining, poor data quality impacting \$3 of every 10\$ of revenue should make it easy for data teams to build a business case around improving data quality.

## Revenue Impact



*What percentage of your company's revenue would you estimate is impacted by having poor data quality?*



**Quick Facts**

Average impact increased 26% to 31% year over year.



**Adam Woods**  
**CEO**  
**Choozle**

"If we have incidents that impeded that visibility, we will hear from customers immediately.

In a previous role, we had a very complex data stack and experienced complications around our data quality, so I have a predisposition to be sensitive to this challenge.

One of my top priorities [was] to identify and fix any issues before they impact customers. We were going to be proactive and buttoned up on this release." [[Full Story](#)].

# Data Trust Is Priceless

The worst type of data monitoring is when it's done by your data consumers. Data trust is everything.

Benefits include increased adoption, faster decision making, and an overall elevation of the data team's role in key initiatives.

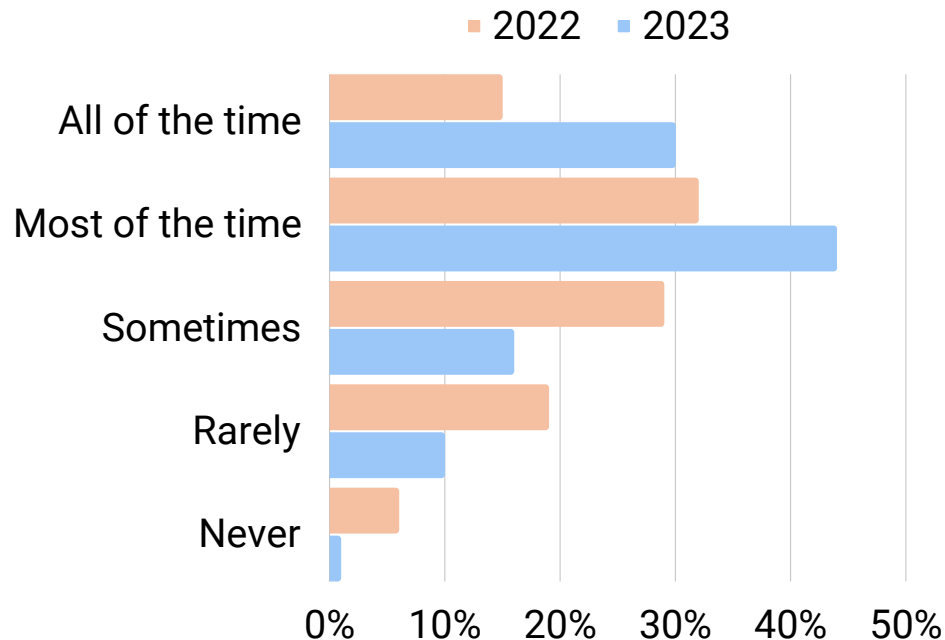
Trust is often assumed until it's lost, usually following a business stakeholder catching a data quality issue first.

Conversely, major improvements in data quality may also go unnoticed, and data trust will be rebuilt slowly following such an incident.

## Business Stakeholder Impact

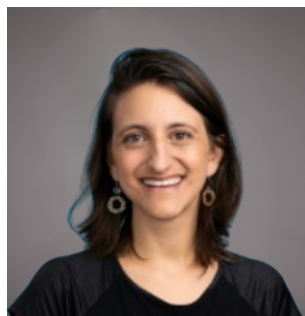


*How often are decision makers or stakeholders impacted by data quality issues that you don't catch?*



**Quick Facts**

**ALL/MOST: Increased from 47% to 74% year over year!**



**Daniel Rimon**  
**Head of Data Engineering Resident**

"Sometimes the CEO of the company would Slack me and my boss and say, 'What's going on?'

Didn't we have any sales?' So that's my nightmare...

[Now] There isn't any process in the company that happens without us.

And people trust and believe our data because it's reliable and it's good. Even if it's complicated, we have the tools to monitor it in real-time and make it more reliable." [\[Full Story\]](#).



# Section 3.

## Ownership & Solutions



# Ownership of Data Quality

## What Structure Is Best?

We've seen every type of ownership structure work, but each has strengths and weaknesses.

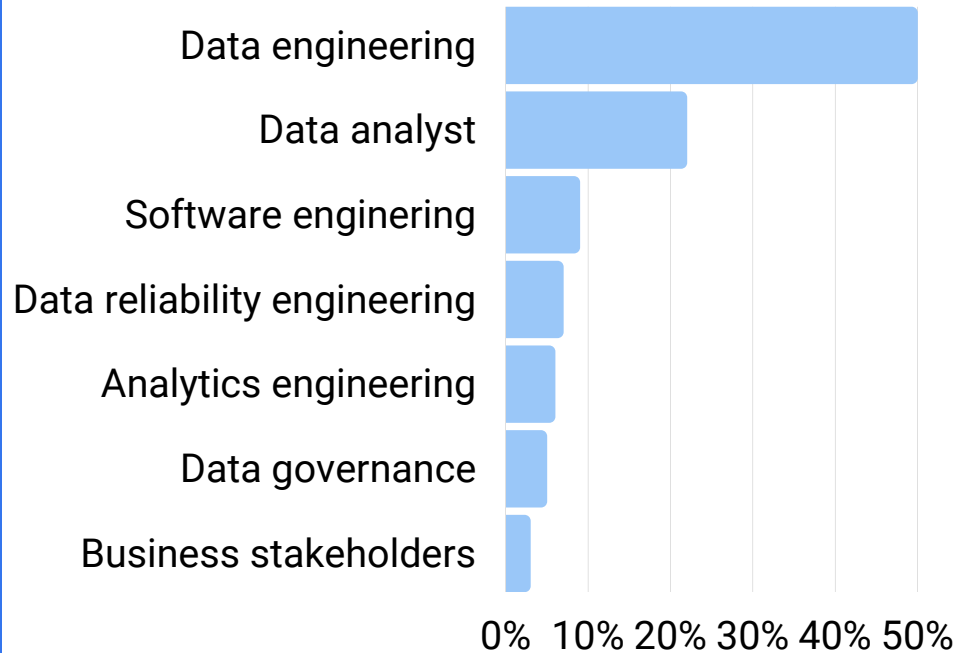
Data engineers are well equipped to solve system-wide problems, but can have limited domain knowledge.

The reverse tends to be the case with data analysts.

Our product telemetry has shown data quality focused teams have positive impacts on operational data quality metrics. It will be interesting to see if this structure gains more adoption over time.



*Who is primarily responsible for data quality at your organization?*



**Quick Facts**

Engineering teams spend more time on data quality.



**Jack Willis**  
**Senior Analytics Engineer**  
**Upside**

“Our analytics engineers are supposed to be accelerants, not necessarily domain experts.

So we position them in the middle of all of these different specialized teams.

This allows them to become a center of excellence and then be able to embed with those teams to acquire the cross-functional expertise

[They] help develop initial solutions, accelerating time to market, and then leave behind the patterns, practices, and teaching for others to solve their problems.” [\[Full Story\]](#).

## Hours To Build ML Detection

### Over-budget Behind-schedule

The data teams we have talked to have scoped out a machine learning data monitoring project as likely to take 8 months requiring work from about 4 full time employees.

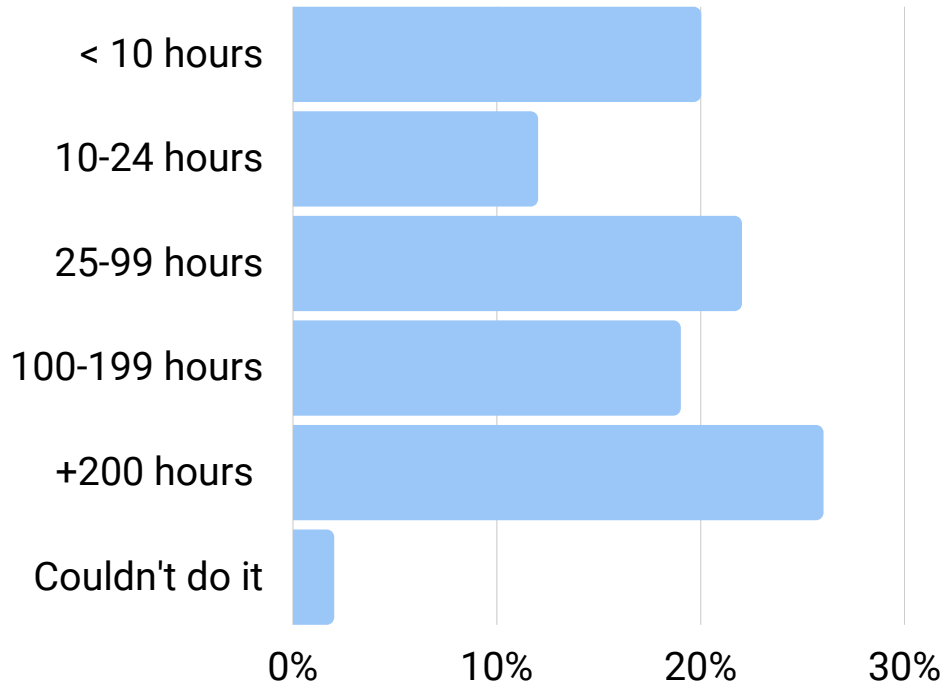
Some of the leading data teams that have actually built similar solutions such as Airbnb or Dropbox have reported similar scopes.

Some survey respondents are perhaps overconfident they could do it in 10 hours or less. But if they can, then they should!

Don't forget after detection, comes root cause analysis and remediation.



*How many hours would it take you to build an automated machine learning driven data monitoring tool?*



*Quick  
Facts*

**AVERAGE- 112 hours**



**Lior Solomon**  
**Former VP Data Engineering**  
**Vimeo**

“We had a schema registry with a fancy framework that knows if the data is right or wrong. But

what it couldn't tell you is if there were any anomalies in the data.

“We use Great Expectations, which is great, but it takes time to implement for the whole pipeline and all the ETLs you have to protect. And unfortunately, usually you prioritize your efforts by how many issues you've had or who is shouting at you more.” [\[Full Story\]](#)

# Why Launch A DQ Initiative?

## A Recognition Of Data's Value

Hopefully this survey has provided some inspiration and urgency for data teams to control their data quality destiny.

If there is one lesson to takeaway it is this: don't wait for the highly visible data incident before implementing a sustainable data quality solution like data observability.

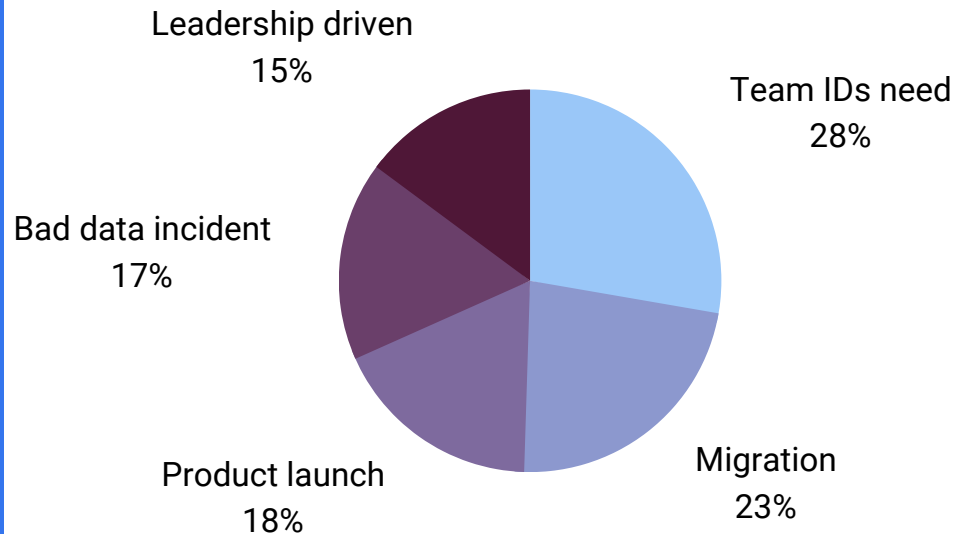
If data is valuable to your company, treat it like such.

Good luck out there.

Reliably Yours,  
Monte Carlo

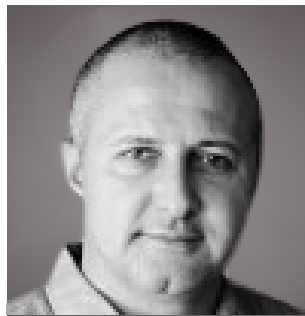


*What is the most likely reason your organization would launch a data quality initiative this year?*



### Quick Facts

Leadership driven correlated to more incidents, longer time to resolution.



**Alex Tverdohleb**  
**VP Data Services**  
**Fox Networks**

“Data observability has become necessity, not a luxury, for us.

As the business has become more and more data-driven, nothing is worse than allowing leadership to make a decision based upon data that you don’t have trust in. That has tremendous costs and repercussions...

It all comes with trust—the moment you drop transparency...people lose trust and it’s really hard to regain it back.” [[Full Story](#)].

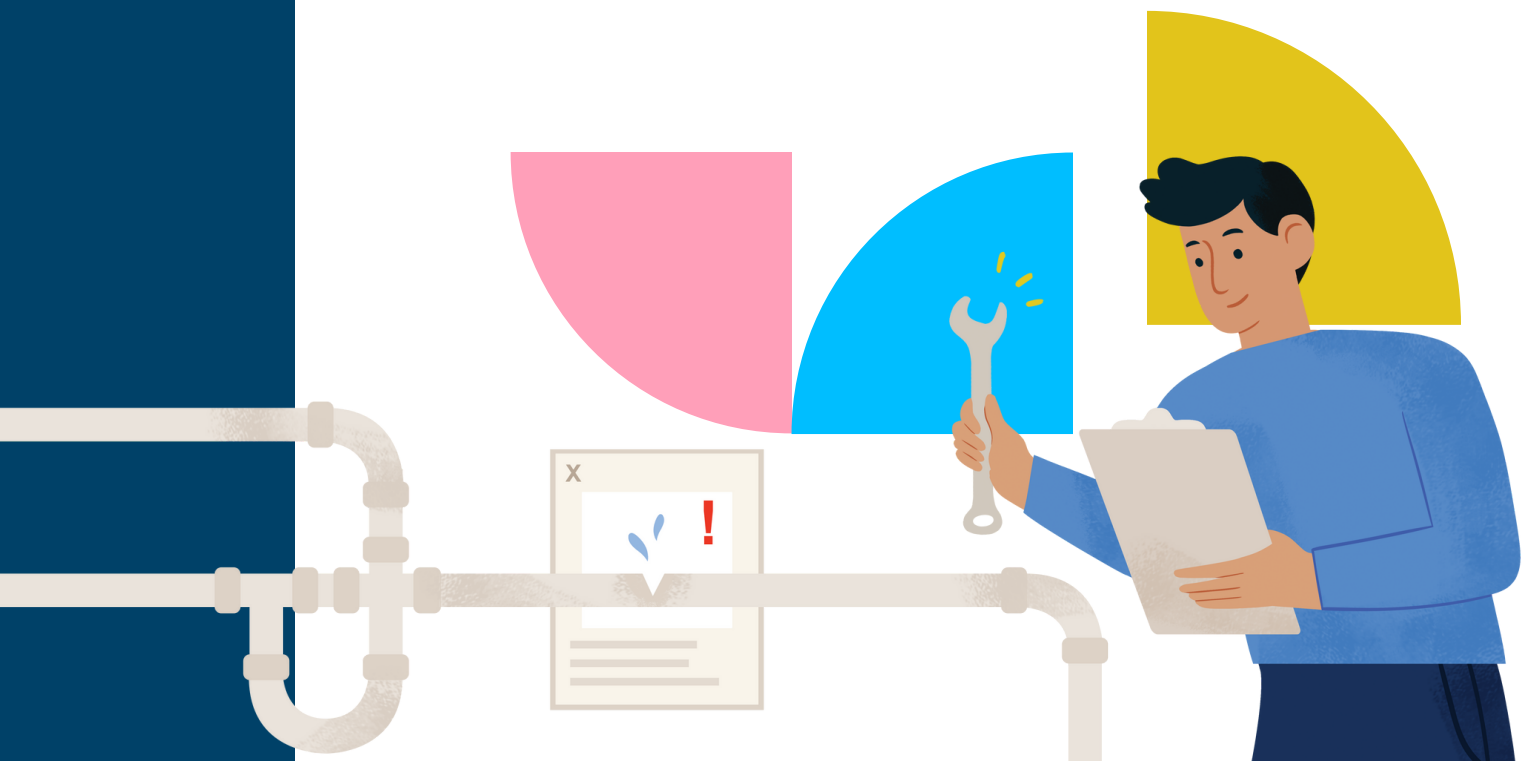
# Section 5. Additional Resources



# Additional Resources

Don't let this be the ending point of your data quality journey! Check out more helpful resources including:

- [Data Downtime Blog](#): Get fresh tips, how-tos, and expert advice on all things data.
- [O'Reilly Data Quality Framework](#): The first several chapters of this practitioner's guide to building more trustworthy pipelines are free to access.
- [Data Observability Product Tour](#): Check out this video tour showing just how a data observability platform works.
- [Request A Demo](#): Talk to our team to get a more accurate assessment of your data downtime, its costs, and what level of value you can expect from Monte Carlo.



WAKEFIELD SURVEY

**2023**

---

**THANK YOU**

LEARN MORE AT  
[WWW.MONTECARLODATA.COM](http://WWW.MONTECARLODATA.COM)



**MONTE CARLO**