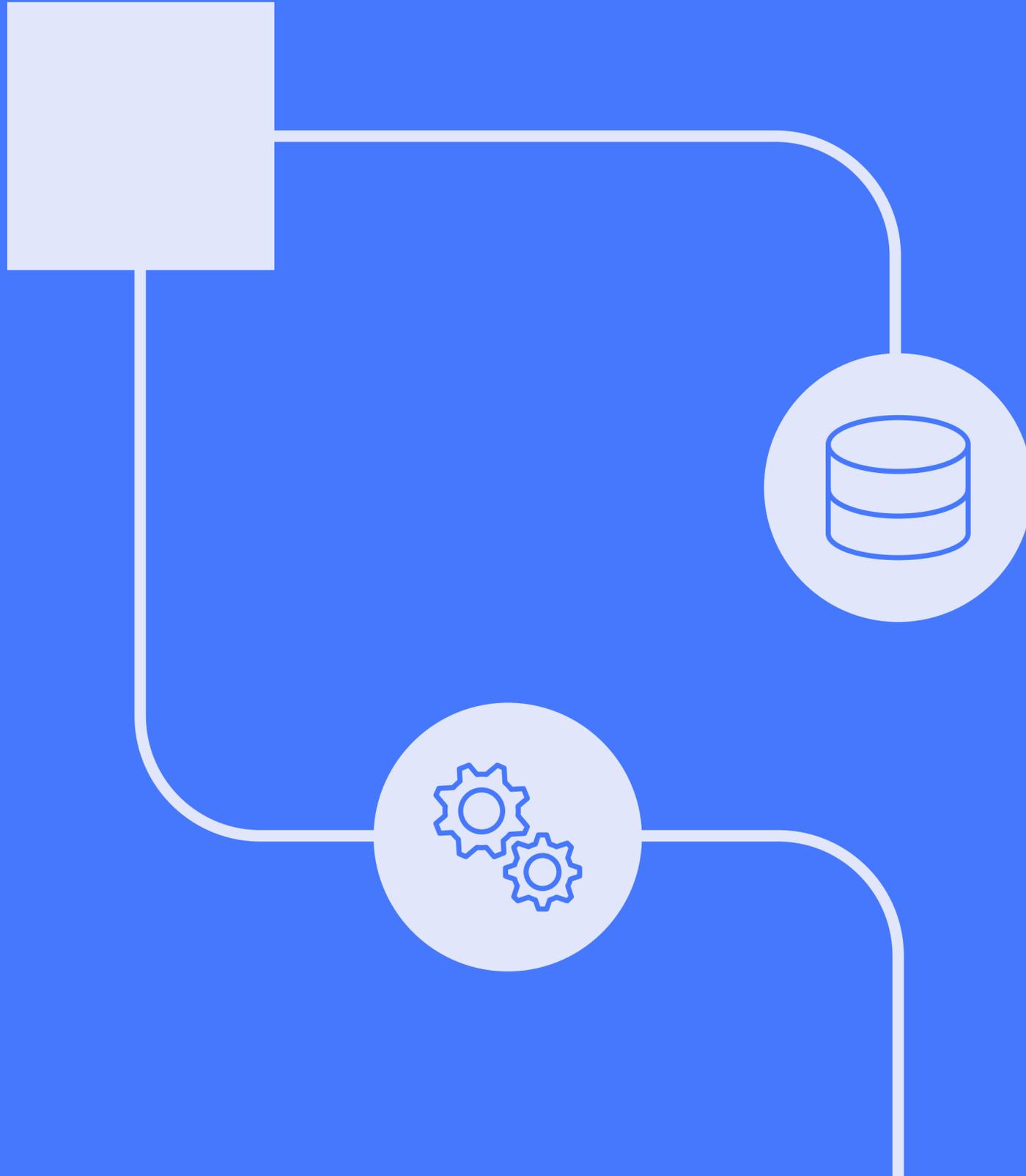




5 Steps to Better Data Quality

For Generative AI & Beyond

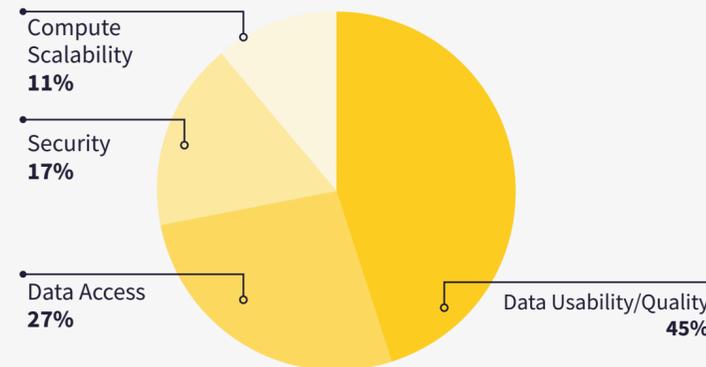


Data Quality Struggles? You're Not Alone.

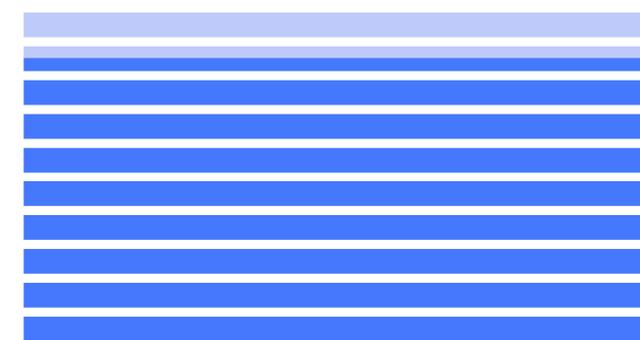
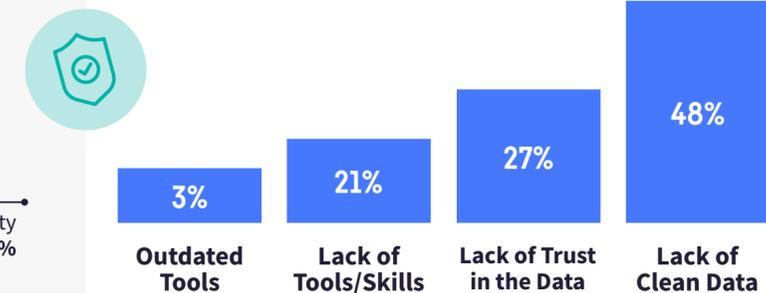
You already know how data quality can impact analytics and so-called “traditional” machine learning (ML) pipelines by causing flawed business decisions or missed opportunities. For example, out-of-date customer information resulting in the wrong products featured for upsell or cross sell, or a spreadsheet with low-quality data leading to erroneous conclusions.

But, as many organizations are currently finding out, data quality also plays a critical role in the success of Generative AI initiatives.

45% of senior analytics and IT leaders cite **data quality and usability** as their **main challenge** with data infrastructure.¹



Of those who cited data quality and usability as their main data infrastructure challenge, nearly half — 48% — said a **lack of clean data** was the greatest **data quality and usability hurdle**.



85% of AI decision-makers believed that **internal data** is high quality and ready for use in AI applications.

However...



¹ <https://pages.dataiku.com/cio-guide-to-modern-analytics>
² Forrester Artificial Intelligence Pulse Survey, September 2023

TYPES OF DATA QUALITY ISSUES



**Unlabeled
data**



**Incomplete or
missing values**



**Data
redundancy**



**Out-of-date or
purely inaccurate
data**



**Poorly labeled
data**



**Inundation of data
sources with no
organization**



**A lack of tools to
properly address
data quality issues**



**Process
bottlenecks**



**Inconsistent or
disorganized data**

Poor data quality can result in large language models (LLMs) learning from incorrect, biased, or incomplete information, which subsequently generates flawed or biased outputs. This can undermine the credibility and effectiveness of Generative AI applications, affecting their adoption and trust among users.

And that's not all: In a world of increasingly stringent AI regulations, data quality matters even more. For example, organizations might need to be able to locate an individual's information quickly — without missing any of the collected data due to inaccuracies or inconsistencies.

Bottom line? The stakes have never been higher.

Data Quality Done Right: A Competitive Advantage

If you're struggling with data quality, you're not alone. But it doesn't make data quality less important. In fact, those who are getting it right are getting ahead.

Before we dive into best practices, it's important to make the distinction between data quality work that happens when ingesting data into, for example, a data warehouse and the work that happens to prepare data for consumption by end users, whether for analytics, ML, or AI.

This compact flipbook will focus on the latter. In particular, we'll look at how to address data quality in a way that allows analytics, data science, ML, AI — whatever the final “product” might be — to flourish.

“



“A lack of quality data is probably the single biggest reason that organizations fail in their data efforts.”

— Jeff McMillan, Chief Data & Analytics Officer, Morgan Stanley Wealth Management

Jeff McMillan shared that the data quality efforts at Morgan Stanley Wealth Management (a Dataiku customer) took about five years to implement in a meaningful way and today make up one of the company's competitive advantages.

”

1. Don't Fall Into the Catch-22 of "Solving" Data Quality



**Go for the right data,
not the perfect data.³**



There is a massive corner of the data and AI software and services world dedicated to data quality. Hundreds of tools and companies promise to address it, and organizations are spending millions to “fix” it.

The problem is these solutions are only capable of addressing part of the problem.

The usual approach is to consider that quality of data must be solved before exposing it to domain owners. But only domain owners have the intricate knowledge of the data required to establish that quality — and they can only address quality by using the data.

Data quality is actually both a requisite and an output of use cases (whether it's analytics, ML, or AI) and, as a consequence, data quality must be integrated in the use case processes themselves and within the interactions with domain experts.

That means companies must start to address data quality first and foremost by accepting it not as a problem that can be solved in and of itself.

³ McKinsey “Moving past gen AI’s honeymoon phase: Seven hard truths for CIOs to get from pilot to scale”

2. Democratize Data Quality

Traditionally, IT owns data quality. The problem is, as previously mentioned, they probably don't know the business data deeply. So as tempting as it may be for IT to wholly own data quality, centralization without a larger goal or purpose won't actually generate business value and, ultimately, will result in data quality efforts falling flat.

That means the problem of data quality isn't always a technological one, but an organizational one that requires synergy across all people and teams. When it comes down to it, many organizations don't have a repository of high-quality and trusted datasets. And, when they do, they may not be accessible in any simple way and available for constant reuse and are, instead, commonly siloed or fragmented.

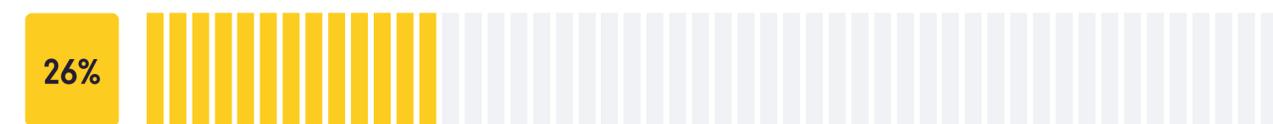
So to avoid the aforementioned Catch-22, organizations must democratize data quality by giving, for example, analysts and business people similar access to data quality understanding as data engineers. In addition to access, there must be a clear and defined process for issue management control. In other words, when people find data quality issues, how do they practically (and quickly) solve them?

Role That Owns the Data Quality Responsibility Within the Organization

- IT/central responsibility



- Decentralized, it's owned by the business



- Ownership is shared, everyone has a role to play



- There is no specific owner





3. Embed Data Quality Across Operations

Giving access to data quality understanding and having a process for issues is the first step, but it doesn't stop there. To improve data quality, you also need to give builders the tools to proactively address data quality in the context of building real use cases.

For example, Dataiku — the Universal AI Platform — offers embedded, as-you-go data quality infrastructure that allows everyone within an organization to have a hand in data quality, ultimately more effectively operationalizing it across the analytics and AI lifecycle. This includes:

- Increased data literacy via shared understanding of data quality status across all stages of the project lifecycle. Everyone working with data — whether for analytics, AI, or Generative AI projects — has an accurate and trusted view.
- Greater control over data quality and the ability to quickly identify and fix issues proactively.
- A unified view of data quality contextualized in data flows, in project dashboards, and alongside universal MLOps views to make data quality a standard view of project health.

Sample settings

Sampling method
First records

Nb. records
10000

Filtering
 Filter the rows of the dataset

Advanced
 Limit memory usage
Max memory usage (MB)
100
 Auto-refresh sample
Automatically refresh the sample when the data changes. The check can be costly for large datasets

UPDATE SAMPLE

state	zip	phone
string US State	string Integer	string Text
Alabama	03404-4384	(308)699-6239x81011
California	01522-1310	(308)699-6239x81011
New York	37768-2026	01142290278
South Carolina	39347-4438	869-985-6299
Kentucky	25794	1-184-809-6600
Connecticut	08086-5571	874-576-0739
North Dakota	15581	(083)182-4626x8012

OK: 49% - NOK: 50% - Empty: <1%

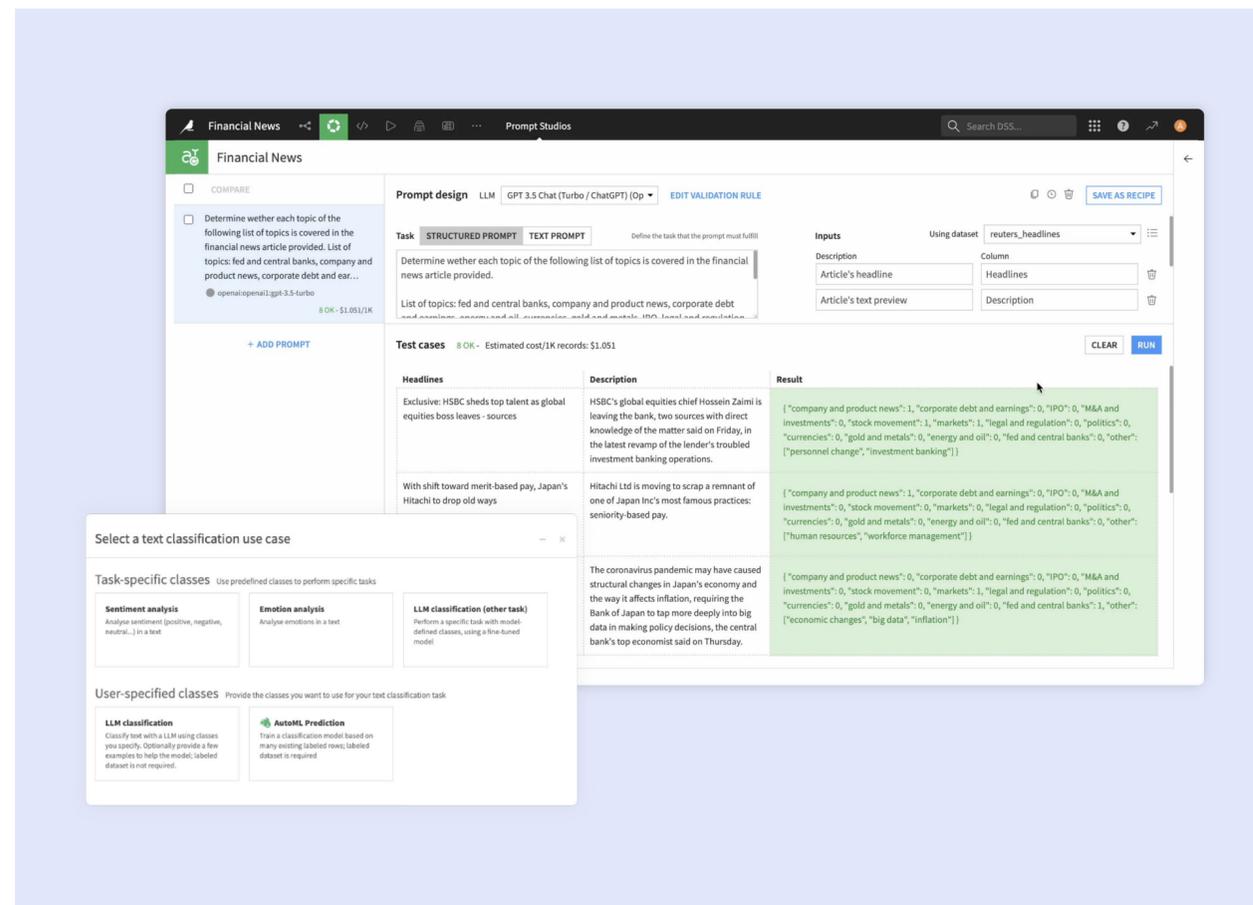
4. Understand How Addressing Data Quality Differs for Generative AI Use Cases

When it comes to Generative AI, data quality still matters — potentially more than ever. Biases or errors present in your real data will inevitably be reflected in Generative AI outputs. Imagine feeding a Generative AI model customer data riddled with inconsistencies. The output might look realistic on the surface, but it wouldn't accurately represent your real customer base.

That said, addressing data quality in Generative AI doesn't look exactly the same as addressing it in a more traditional data pipeline. For example, it might look like this:

- Creating a Retrieval Augmented Generation (RAG) system and training on high-quality inputs.
- Model fine-tuning to supplement models and reduce hallucinations.
- Prompt engineering to identify the best-performing prompt/model combinations that result in the highest level of data quality and accuracy.

The advantage of using a platform like Dataiku is that you don't have to buy another tool, or switch tools, to address low-quality Generative AI responses. Builders still have the power to take data quality issues into their own hands and address them as they're shaping those use cases.



5. Make Data Quality Part of Larger Governance Efforts

The final step to getting ahead with data quality is widening the thinking around data quality from a narrow problem to part of larger governance efforts around data and AI projects.



This might include aligning on a clear definition of what “quality” means. For example, different lines of businesses using the same data (e.g., marketing and finance both use customer data) may have different standards and, therefore, different requirements and expectations for data quality initiatives.

It might also include establishing a process for the ongoing monitoring of data quality metrics and implementing continuous improvement initiatives to address emerging issues and maintain high standards. For example, if data shifts over time and is causing an ML or Generative AI model in production to behave poorly, how does this problem get identified and who addresses it? AI governance, operations (whether MLOps or LLMOps), and data quality are — at the end of the day — all intertwined.

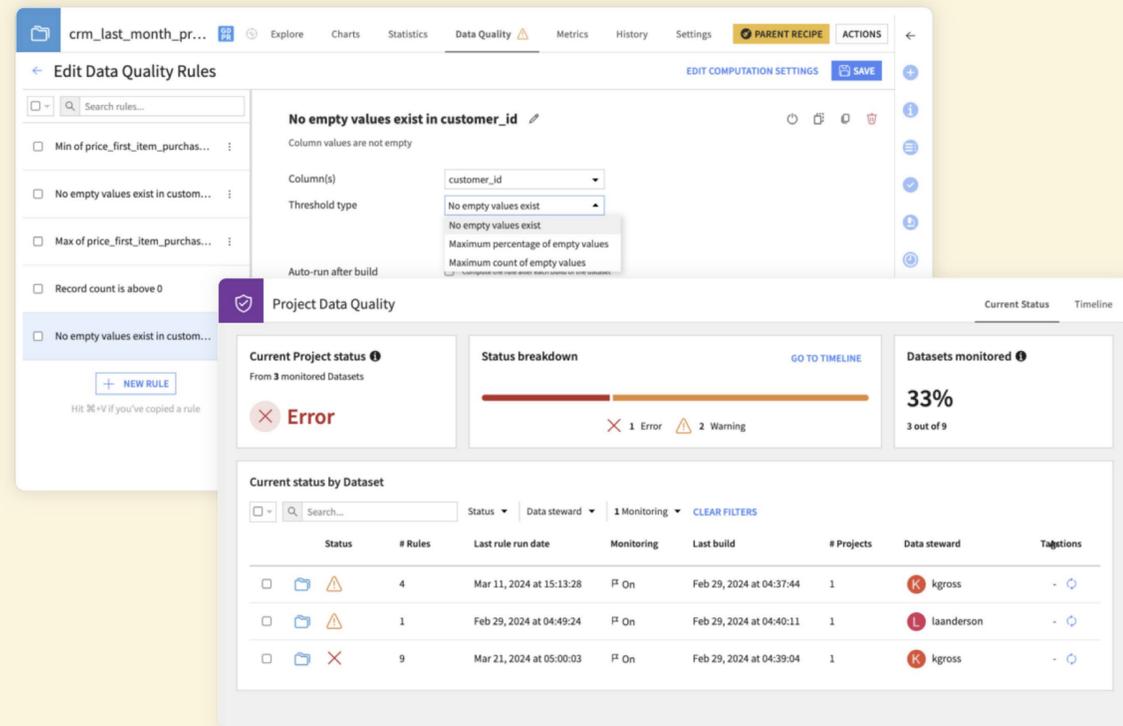
Bonus:

Data Quality in Action With Dataiku

Dataiku customer Bankers' Bank uses Dataiku to ensure data quality across an array of financial analytics. As a result, the team has been able to reduce the time to prepare analyses and deploy insights by 87%.

They do transactional reporting on the different volumes of transactions they process, which was previously very manual. Data was pulled in from various sources (including a CRM) and validation included extensive backtracking to pinpoint where any errors occurred without visibility to the entire data flow, which was — at times — nearly impossible.

Thanks to their use of Dataiku, the Universal AI Platform, the team has been able to reduce time associated with pulling that data while simultaneously improving the data quality and reliability.



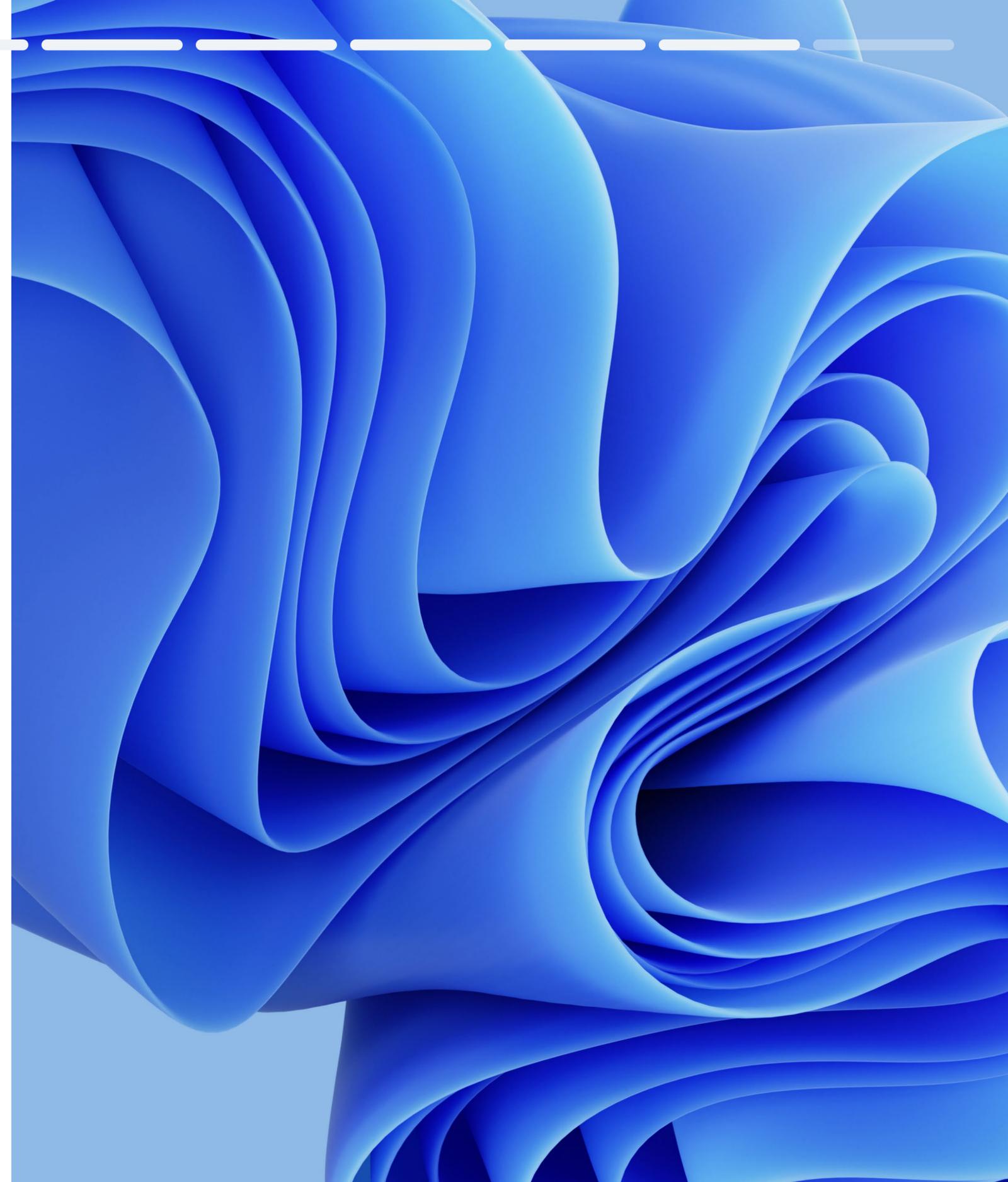
The screenshot displays the Dataiku Data Quality interface. The top section, 'Edit Data Quality Rules', shows a rule configuration for 'No empty values exist in customer_id'. The rule is set to monitor the 'customer_id' column with a 'No empty values exist' threshold type. Below this, the 'Project Data Quality' dashboard provides a 'Current Project status' overview, showing a 'Status breakdown' with 1 Error and 2 Warnings, and 'Datasets monitored' at 33% (3 out of 9). A table titled 'Current status by Dataset' lists the following data:

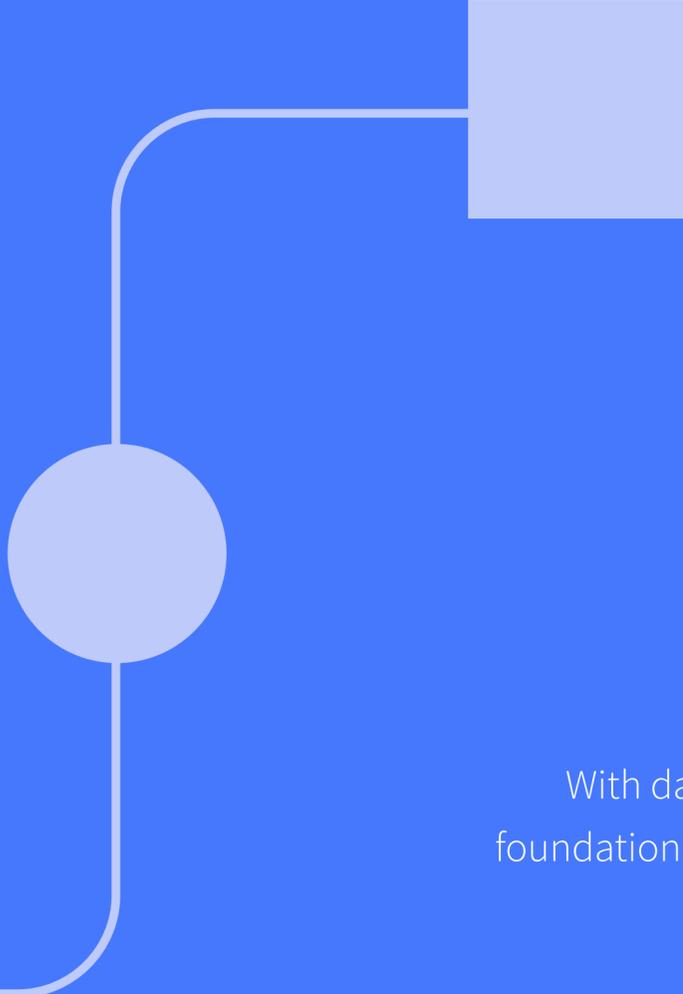
Status	# Rules	Last rule run date	Monitoring	Last build	# Projects	Data steward	Tagetions
⚠	4	Mar 11, 2024 at 15:13:28	🟢 On	Feb 29, 2024 at 04:37:44	1	kgross	-
⚠	1	Feb 29, 2024 at 04:49:24	🟢 On	Feb 29, 2024 at 04:40:11	1	laanderson	-
✖	9	Mar 21, 2024 at 05:00:03	🟢 On	Feb 29, 2024 at 04:39:04	1	kgross	-

Final Thoughts

Effective data quality management is key for delivering accurate, reliable insights that drive strategic decision-making. By thinking of data quality as ongoing work as part of building use cases, organizations can unlock the full potential of their data assets and gain a competitive edge in today's AI-driven landscape.

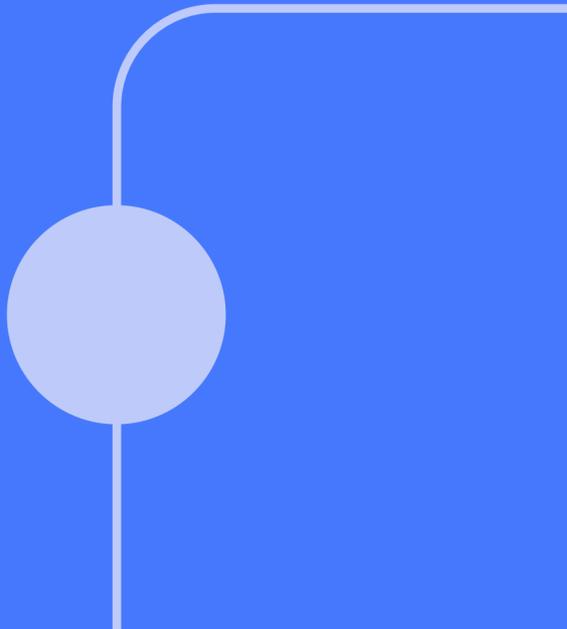
That means data quality management is not a turnkey initiative that is handled all at once. Rather, it's an ongoing process that needs to involve the business from the beginning in order to be successful.





Improve Your Data Quality With Dataiku

With data quality features in Dataiku, have confidence that your insights are built on a solid foundation. Track, verify, and fix data quality so that you can deliver powerful (and trusted) insights.



LET'S GO