great expectations

State of Data Quality

REPORT

2022

Table of Contents

Introduction

With our State of Data Quality survey, we set out to identify and quantify the major pain points that data practitioners encounter, as well as the consequences of poor data quality within organizations.

Data practitioners continue to face a rocky data quality landscape. Data quality issues remain visible problems, and a large number of data practitioners do not have the tools or resources to address those problems.

Data practitioners identified the effects of poor data quality as manifesting in diverse ways. Production and product launch delays (27%), teams and individuals creating their own data stores (27%), and data scientists spending too much time preparing data for analysts (26%) were the most commonly cited instantiations of low data quality.

Just over half of data practitioners (51%) reported having active data quality processes in place, with 75% actively carrying out data validation. They felt that leadership was generally supportive of data quality efforts, with 89% reporting leadership that was very or mostly supportive.

However, less than half of data practitioners (49%) reported high trust in their data, with 13% having low trust.

We also asked respondents about the composition of their data stack, the languages they use, and their major data initiative plans for the near future.

Methodology

The survey was conducted in May 2022 by Pollfish, an independent research platform. Responses were gathered from 500 information services and data professionals based in the United States.

Respondents were aged 18-54, with 57% identifying as men and 43% as women. 60% of respondents were employed at companies with 250 or more employees.

The Pollfish survey was supplemented by information from the Great Expectations Slack community.

Analysis

Summary

Our analysis revealed that data quality continues to be a pervasive issue, despite increasing awareness of its significance. 91% of respondents said that data quality had some level of impact on their organization, but only 23% characterized data quality as part of their organizational ethos.

It remains difficult to generalize data quality problems except in the broadest possible terms. The driving factors behind data quality issues all fall under the umbrella of the idea that teams within organizations are poorly aligned and face significant gaps in both understanding and technical capabilities.

Symptoms of data quality are visible throughout the data lifecycle, with some factors such as production delays cited more often than others but no singular symptom holding a commanding lead.

Roughly half of data practitioners indicated that their organizations had no active data quality, including 41% of respondents with high trust in their dataset; 28% of organizations overall either do not think they need data quality or have prioritized other things, including 27% of organizations with high trust.

The proportion of respondents with low trust in their data was also small (13%), further indicating a widespread disconnect between data practitioners' perception of the ubiquity of data problems and their confidence in their datasets as a whole.

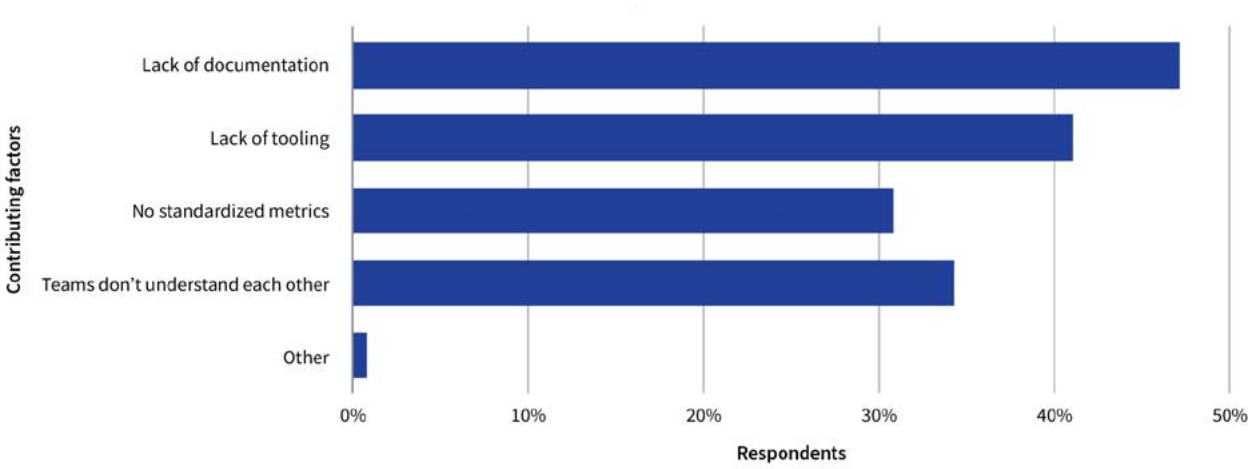
But while the overall state of data quality within organizations remains unsettled, data practitioners indicated active and widespread engagement with both data quality and their data usage as a whole. 75% of respondents are currently validating their data, and nearly half of the organizations with no active data quality have concrete plans to implement it.

Maintaining and increasing this commitment to data quality will be key; our results indicated that data landscapes are far from settled. Respondents were planning an average of 2.5 new initiatives each within the next quarter; for 29% of data practitioners, this included new data sources.

What do you see as contributing factors to data quality issues at your organization?

The top-reported factor contributing to data quality issues, selected by nearly half of respondents (47%), is a **lack of documentation**. Almost as common was a **lack of tooling** (41%).

34% of respondents indicated a lack of unified understanding because **teams don't understand each other** as a concern. 30% specifically identified that there were **no standardized metrics.**

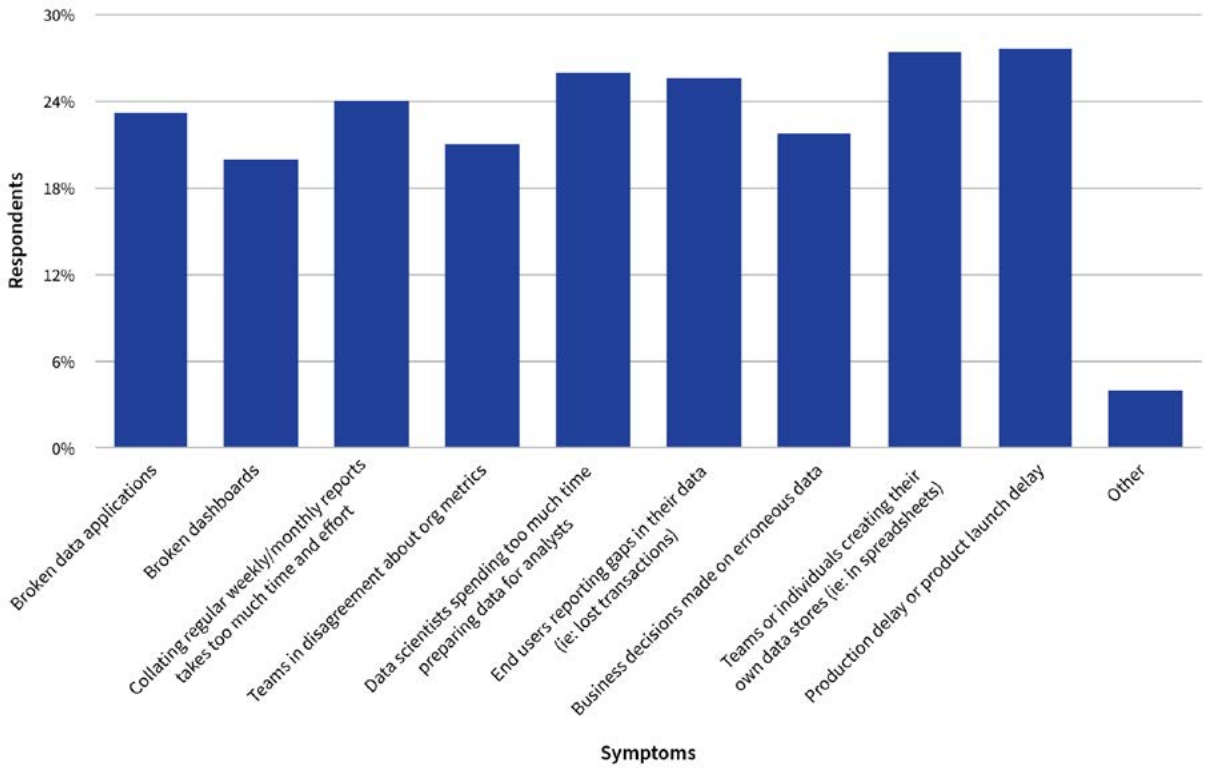


Overall, these results indicate a significant need for a shared standard of data quality to support understanding throughout the organization, as well as a data quality strategy that can produce readable documentation.

What are the top symptoms of poor data quality you see at your organization?

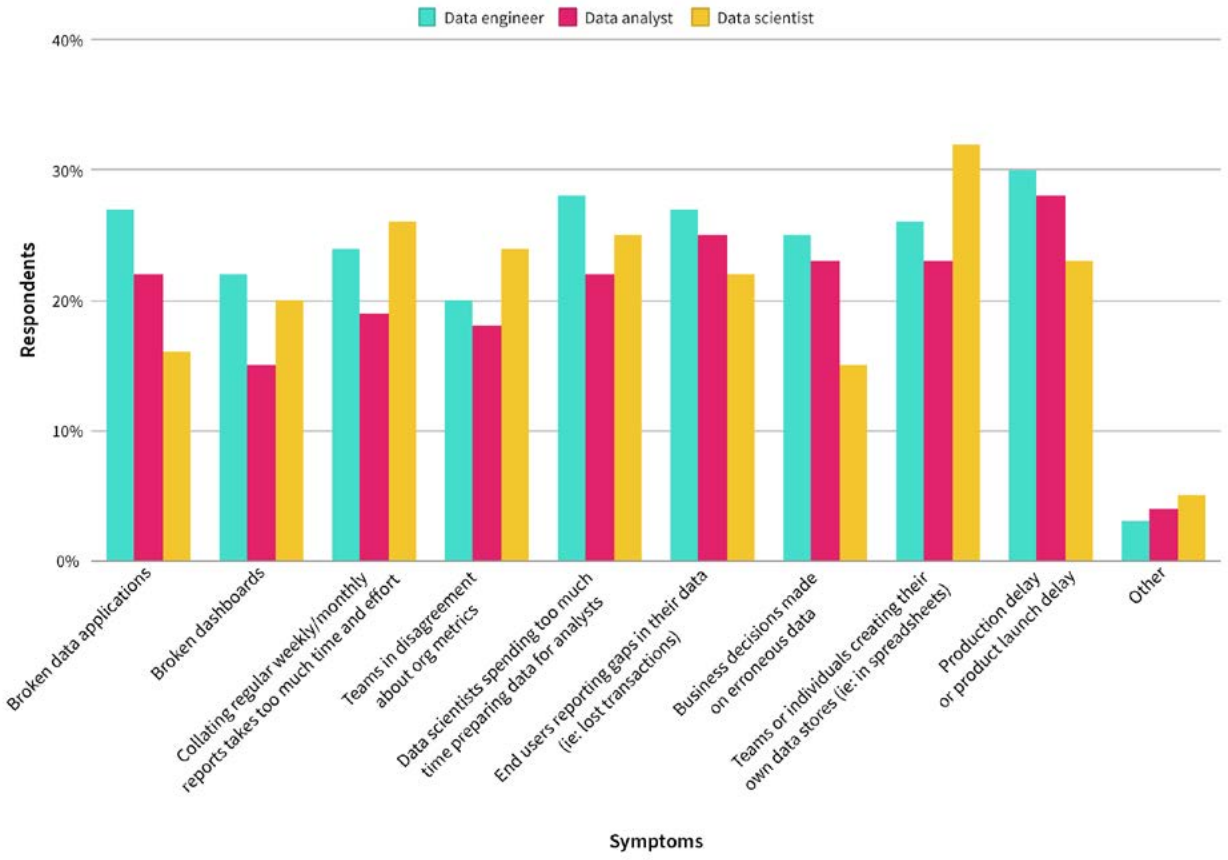
Respondents attributed a wide array of problems across the data lifecycle to DQ issues, including problems with data creation, analysis, and reporting.

Although **data scientists spending too much time on data preparation** is practically a data quality cliche by now, it's still relevant: more than a quarter of respondents (26%) cited this as a top symptom.



Data scientists themselves found **teams or individuals creating their own data stores** to be most troublesome, with 35% of them identifying this as an issue.

For data engineers and analysts, the ripple effect of data quality challenges was most pressing. **Production delay or product launch delay** was the top symptom, concerning 30% of data engineers and 28% of analysts.



In your own words: What is the biggest data quality challenge at your organization?

Multiple silos from M&A acquisition.

Data engineer

Our data is pretty clean, but when our pipelines fail, they can fail in obscure ways.

Data engineer

Analytics is treated as nonproduction so there are no SLAs and pipelines have many hidden, brittle dependencies.

Data scientist

Inconsistencies in data due to some key metrics relying on manual entry (e.g. manually activating user subscriptions etc.), inconsistencies in 3rd party integrations/inability to access raw data (Stripe), production code changes that affect metrics but aren't proactively socialized with the data team.

Data scientist

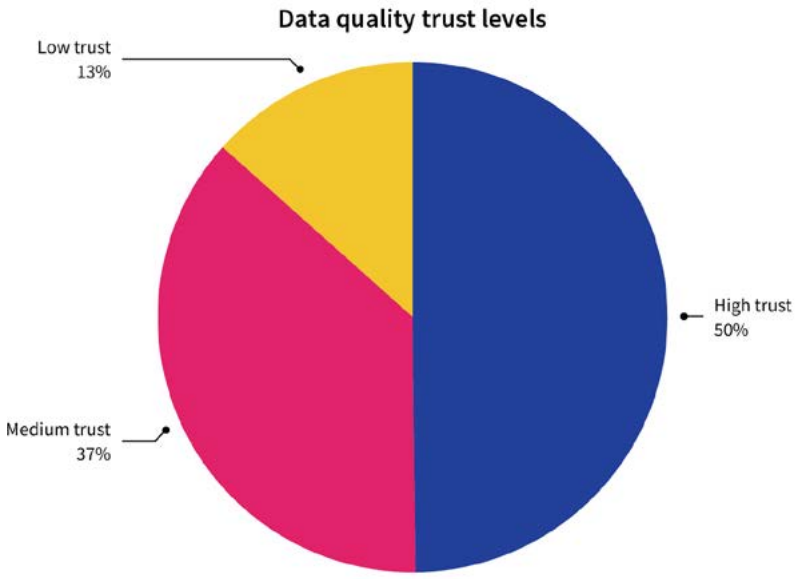
Implementing a consistent data quality framework with high observability and low levels of manual configuration.

Data engineer

How much do you trust your data?

We asked data practitioners about their trust in their data, ranking it as **high trust, medium trust,** or **low trust**.

In the aggregate, just under half of data practitioners reported **high trust** in their data.



Dividing respondents into upstream and downstream users uncovered a starker divide. 55% of engineers and 50% of data scientists had high trust in the data, but only 34% of analysts felt the same. We theorize that this split reveals that the effects of poor data quality generally are visible on the business side before they become a significant technical issue.

Data practitioners who had **medium trust** followed an inverse pattern. While 46% of data analysts had medium trust in their data, 35% of data engineers and 33% of data scientists felt the same.

Respondents with **low trust** were more evenly distributed across the data stream. 13% of respondents overall had low trust in their data, including 9% of data engineers, 20% of data analysts, and 16% of data scientists.

HIGH TRUST

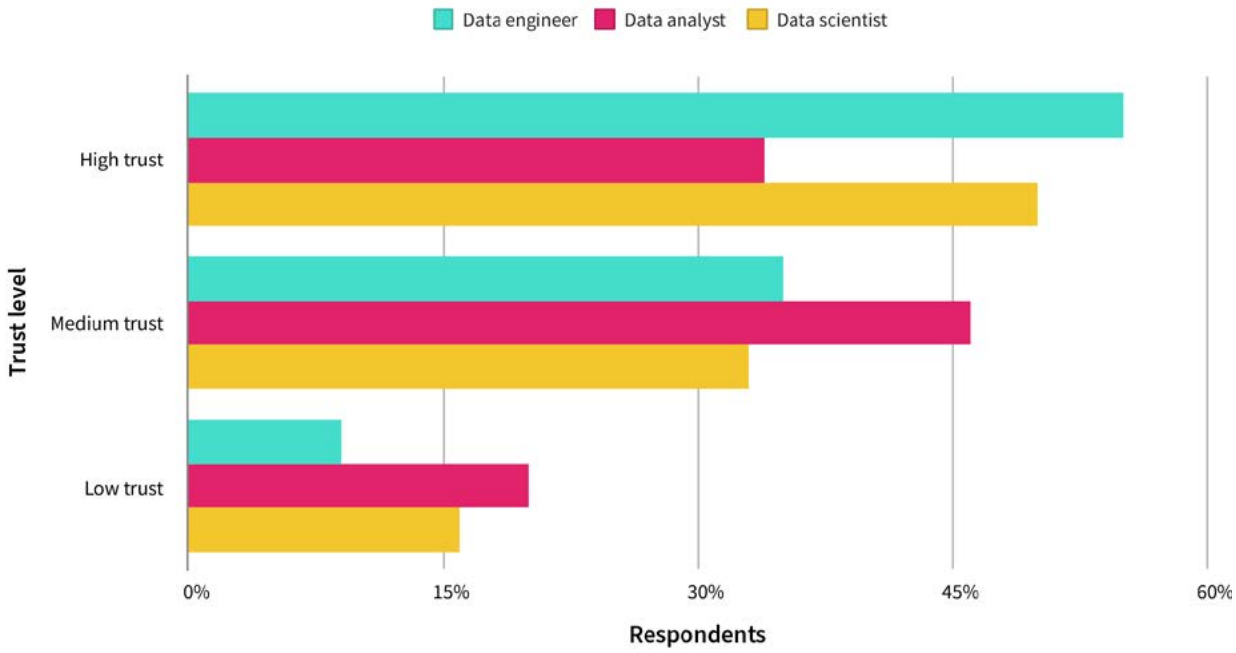
Downstream application performance, confidence in business decisions informed by good data; close understanding about metrics and how they perform; little or no friction with teams.

MEDIUM TRUST

Applications work with some manual fixes; some alert fatigue, some data-driven business decisions; alignment on some metrics; a lot of friction between teams.

LOW TRUST

Broken apps/dashboard, many decisions made on unreliable/bad data; teams have no shared understanding for metrics; siloed or conflicting departments.



The total proportion of respondents who indicated **high trust** in their data was larger than we expected given the other issues they reported. Paired with the trust divide between upstream and downstream data users, this reveals that contributing factors to a high-trust perception despite across-the-board acknowledgment of data quality problems may include:

- The difficulty of gauging the effects of any one data quality incident on the organization's data as a whole.
- Data silos limiting practitioners' knowledge of DQ issues to only a segment of the company's data, with the unknown data considered high-quality in the absence of evidence otherwise.
- Role-based silos making it difficult for respondents to identify when data quality incidents occur outside of the scope of their role.

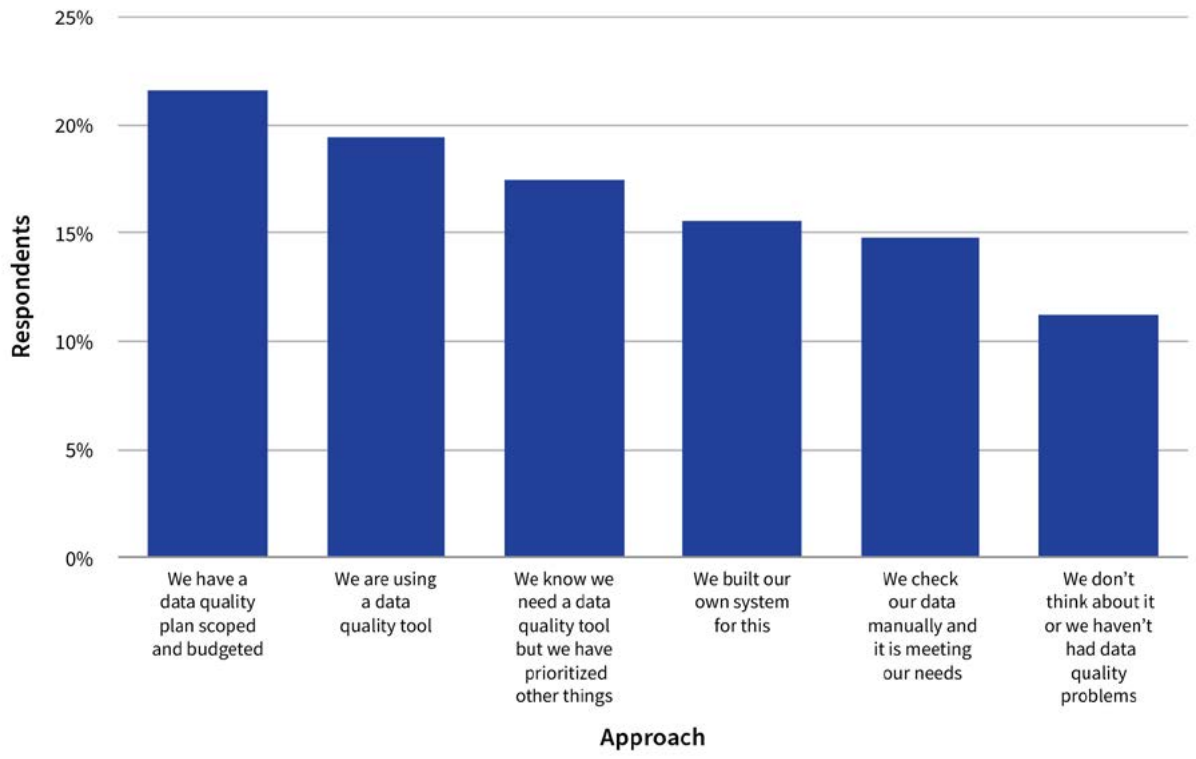
Better communication between teams and roles could alleviate most of these issues, providing an avenue not just for improved cross-team/cross-role understanding of data quality, but opening a pathway to additional perspectives for solving those data quality issues.

What's your approach to data quality?

Respondents were asked to choose a single option that best described their organization's current approach to data quality. Results revealed a roughly 50/50 split between organizations that are actively engaging in a data quality process and organizations that are not.

The most-selected option, picked by 21% of respondents, indicated that **we have a data quality plan scoped and budgeted**, intimating future data quality plans. At the other extreme, 11% of data practitioners chose **we don't think about it or we haven't had data quality problems**.

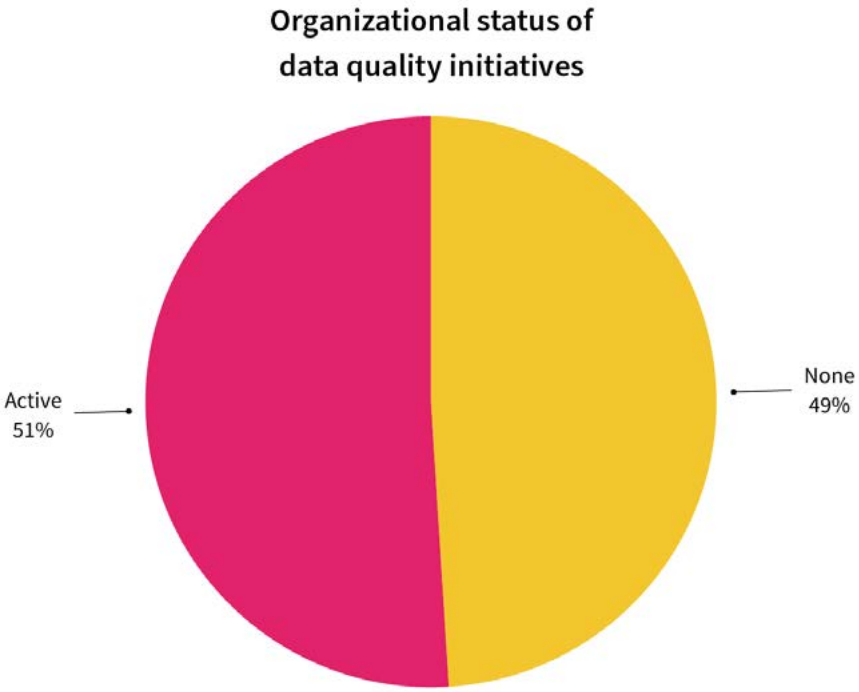
Another 17% of respondents selected **we know we need a data quality tool but we have prioritized other things**.

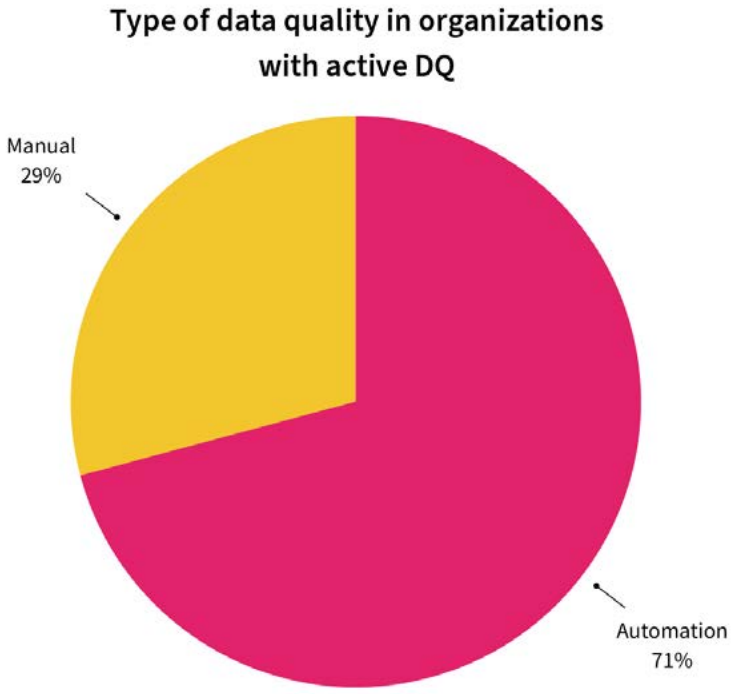


Put together, 49% of data practitioners have **suboptimal data quality or no data quality at all** in their organization.

On the flip side, 14% of respondents chose **we check our data manually and it is meeting our needs.**

Unsurprisingly, most respondents with active data quality used automation: 19% of data practitioners indicated that **we are using a data quality tool**, and 15% said that **we built our own system for this**.



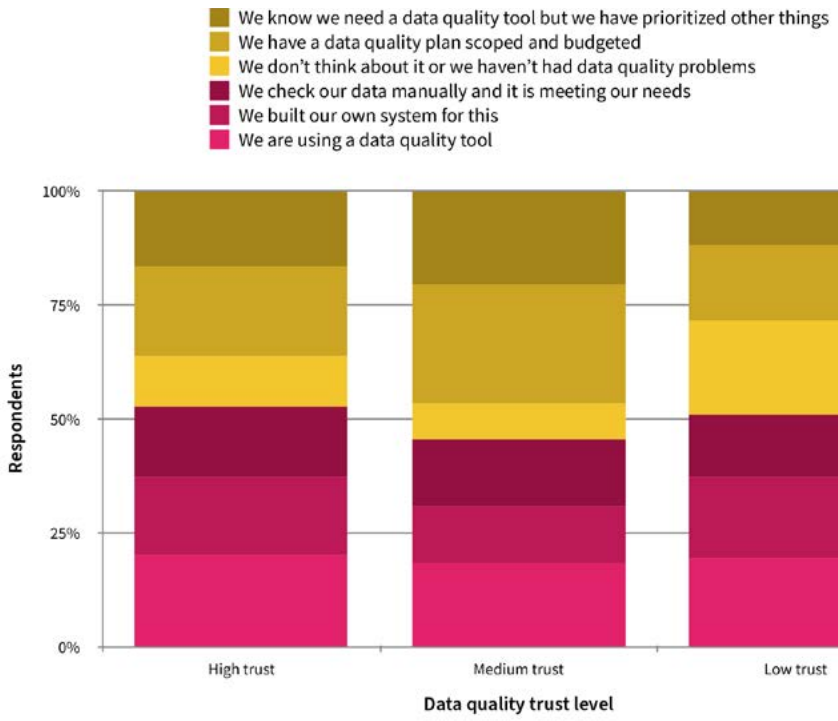


While the theoretical conversation around data quality is fairly well-developed, our results show that actual data quality implementation is moving much more slowly. This is unsurprising and can be explained by a number of factors, including:

- General organizational inertia.
- The difficulty of implementing new technology around production data.
- Lack of a shared, open standard for teams to easily start with, among other factors.

More interestingly, an organization's current approach to data quality is not significantly correlated with its level of trust in the data. Respondents with **low trust** responded that they **don't think about or haven't had** data quality problems in noticeably larger proportions than respondents with high or medium trust, and **medium trust** respondents were the group relatively most likely to say that their plan was **scoped and budgeted** or they **know we need a data quality tool but have prioritized other things**.

But overall, the roughly half-and-half split between **active data quality** and **no active data quality** was maintained across data trust levels.



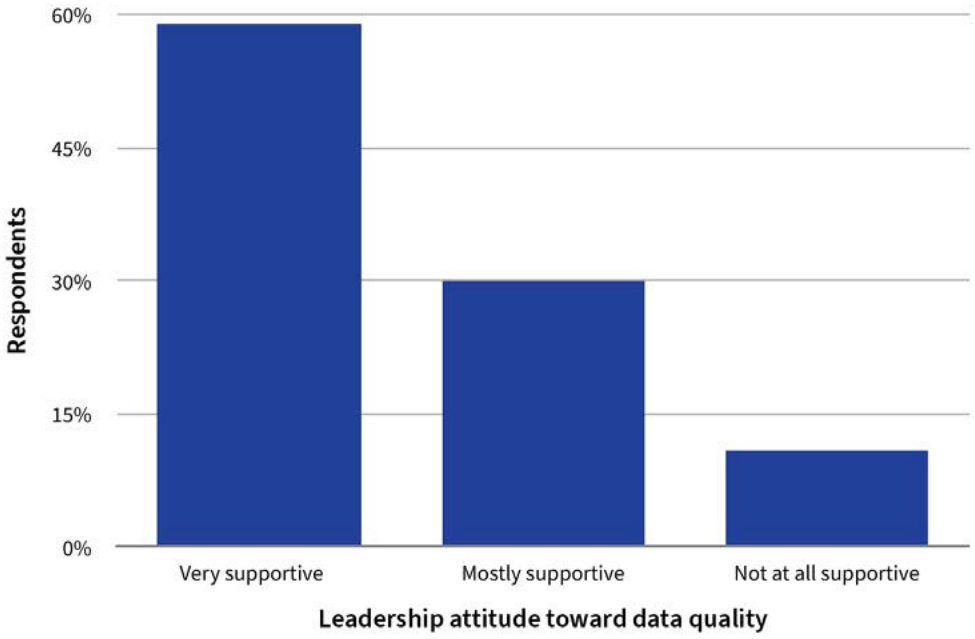
We theorize that this is due, at least in part, to suboptimal capabilities in data quality measurement. Without widespread and standardized ways of quantifying the organization's level of data quality, data practitioners have no choice but to rely on fragmentary evidence (at best) and their gut feeling (at worst) when estimating their organizational data's trustworthiness, and default to assuming that data they are not familiar with is high-quality.

How supportive of data quality is leadership?

We asked data practitioners for their impression of how supportive their organization's leadership was of data quality efforts.

The vast majority reported supportive leadership, with 59% of respondents indicating that leadership was **very supportive** regarding data quality, and an additional 30% of data practitioners having **mostly supportive** leadership.

Only 11% of respondents had leadership that was **not at all supportive**.

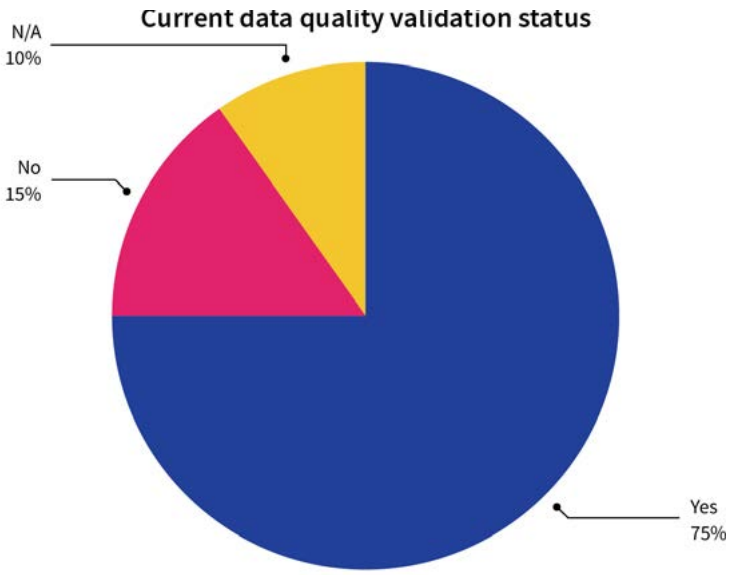


The responses to this question indicate widespread acknowledgement of the importance of data quality at high levels throughout the organizational hierarchy.

They also demonstrate how support alone is not enough to bring data quality to fruition; recognition of its importance is not the primary barrier to widespread data quality implementation.

Are you doing data validation right now?

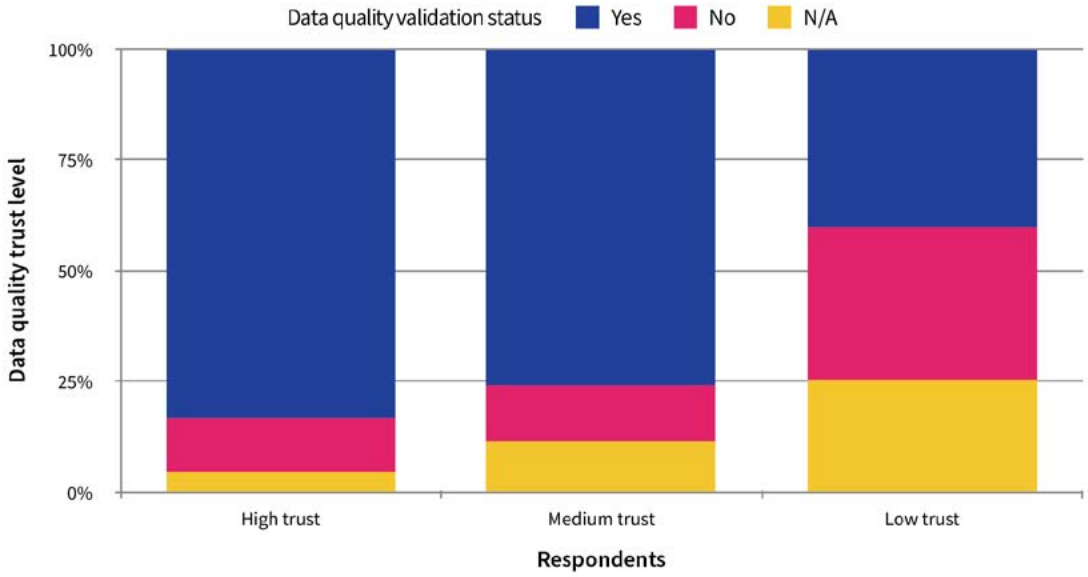
Respondents were asked if they are currently validating their data. 75% of data practitioners responded **yes**; 15% said **no**, with around 9% indicating that this question was not applicable.



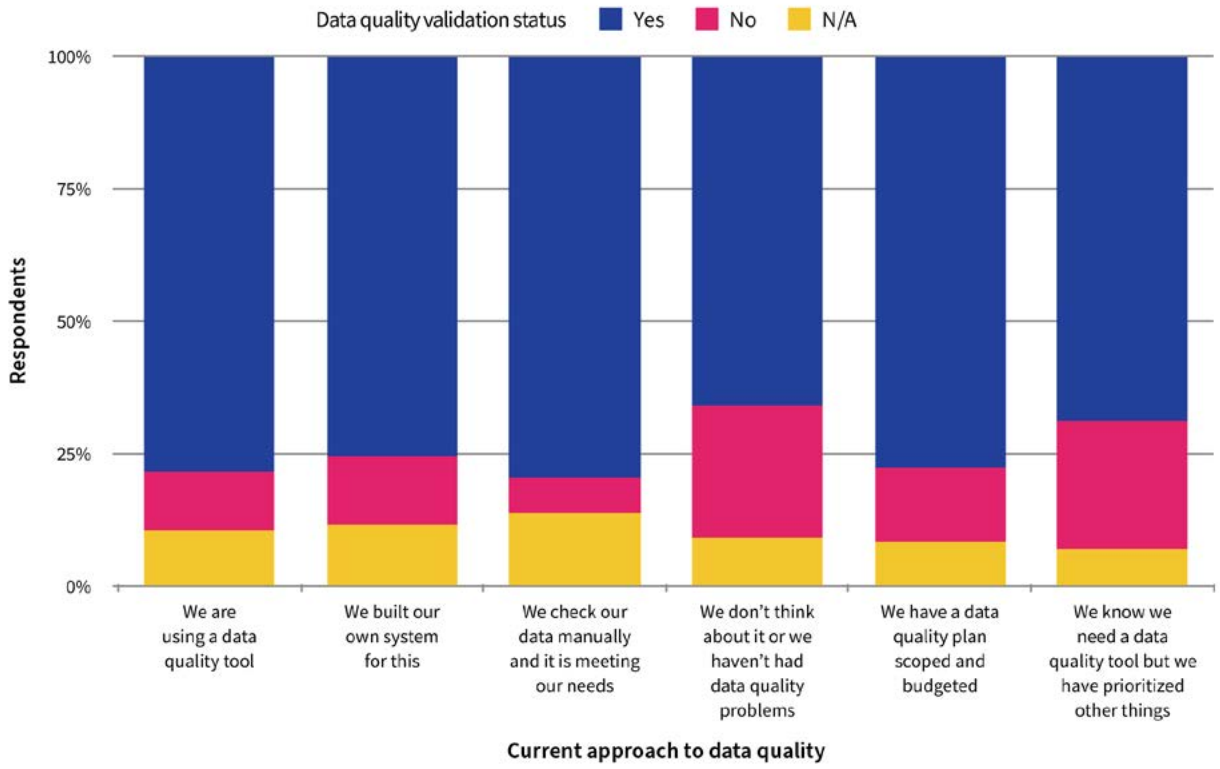
Validation is one of the data quality activities for which it is relatively easy to obtain SaaS validation, particularly for widespread data types like addresses and contact information.

Given that, it's unsurprising that the proportion of data practitioners who have active data validation is significantly higher than the proportion who consider themselves actively engaging in data quality.

Doing data validation is a good predictor of a data practitioner having medium or high trust levels in the data.



However, data validation is not useful for predicting the respondent's approach to data quality in any meaningful way. While respondents who do not have active data quality generally report lower rates of data validation than respondents with active data quality, it is not a significant difference.



We hypothesize that this is attributable to one or both of the following:

- Respondents do not consider data validation to be an aspect of data quality. We see this as a potential cause particularly if the main provider of data validation is a SaaS service that is marketing itself specifically as a validation and not attempting to position itself within the wider umbrella of data quality.
- Respondents perform or are aware of data validation that is performed in specific scenarios or to specific datasets, but these validation efforts are not widespread or coordinated at the organizational level.

In your own words: If you could change one thing about how data quality is handled at your organization, what would it be?

Higher priority for dev teams to provide necessary information.

Data scientist

Good GUI to collaborate and manage data quality tests.

Data engineer

Handle it first, not an afterthought.

Data engineer

Better tooling for automated checks.

Data scientist

Incentivise feature store integrity and re-use.

Data scientist

Make them be more modular and flexible, so we can construct our data validation rules easier.

Data engineer

Defining a consistent, interoperable, open standard for data quality.

Data engineer

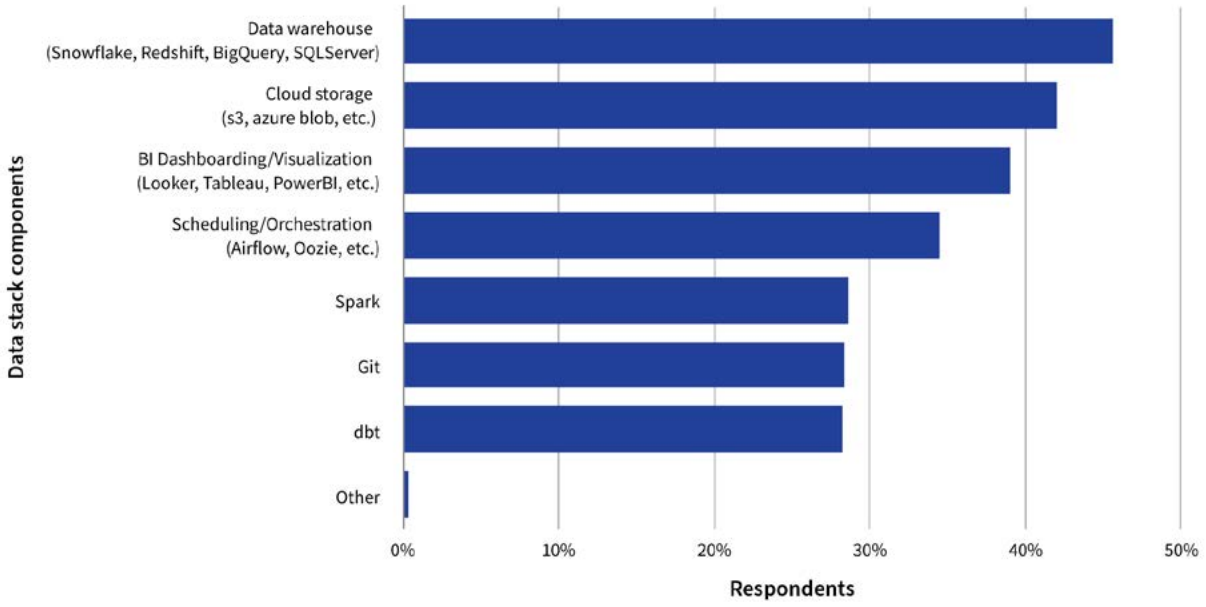
What components are in your data stack?

We asked data practitioners to indicate what components their data stack included.

Most common were **data warehouses**, used by 45% of data practitioners; **cloud storage**, in use by 42% of respondents; and **BI dashboarding/visualization**, present in 39% of data practitioners' stacks.

Lagging only slightly behind these major players were **scheduling/orchestration** components, used by 34% of respondents.

Spark, **dbt**, and **Git** were each in use by 28% of data practitioners.



As expected, responses showed a wide range of stack compositions. This clearly illustrates the need for data quality solutions that have a wide range of integration capabilities and can be flexibly configured to fit unique pipelines.

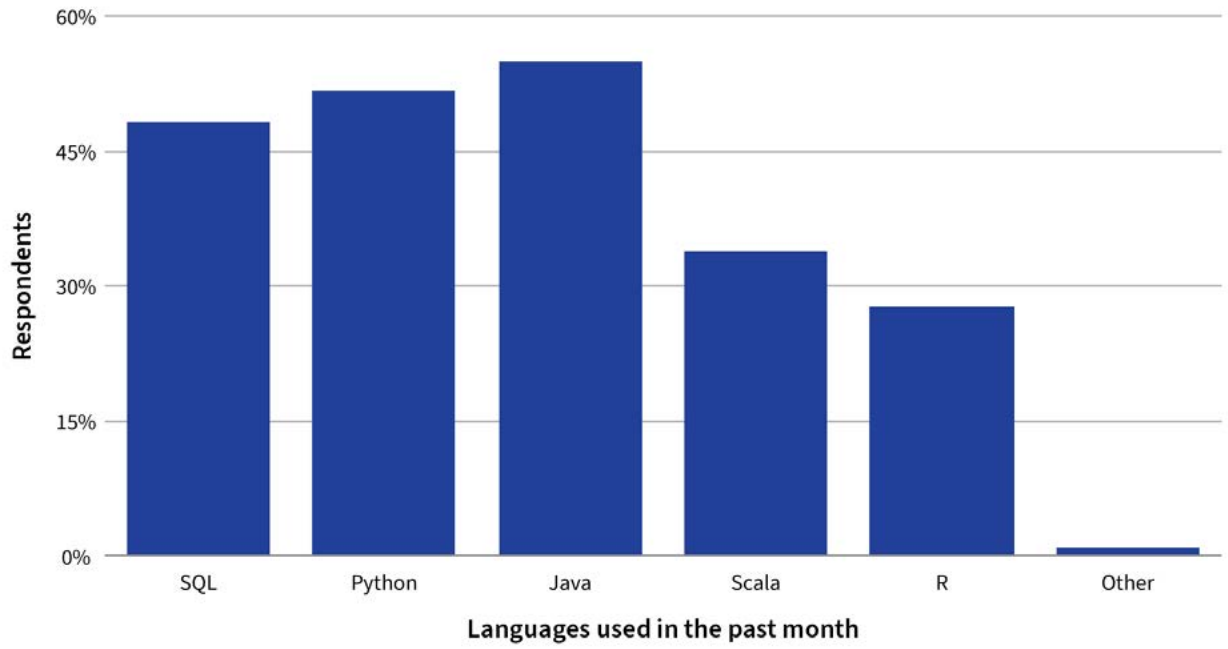
Only 0.4% of respondents indicated that there was an **other** component in their stack.

While this number is undoubtedly higher in reality, its drastic difference from the other answers shows that—for now—there is a core set of components that data quality services can focus on to serve the vast majority of data practitioners.

What languages do you use?

We asked respondents what languages they had used in the last month.

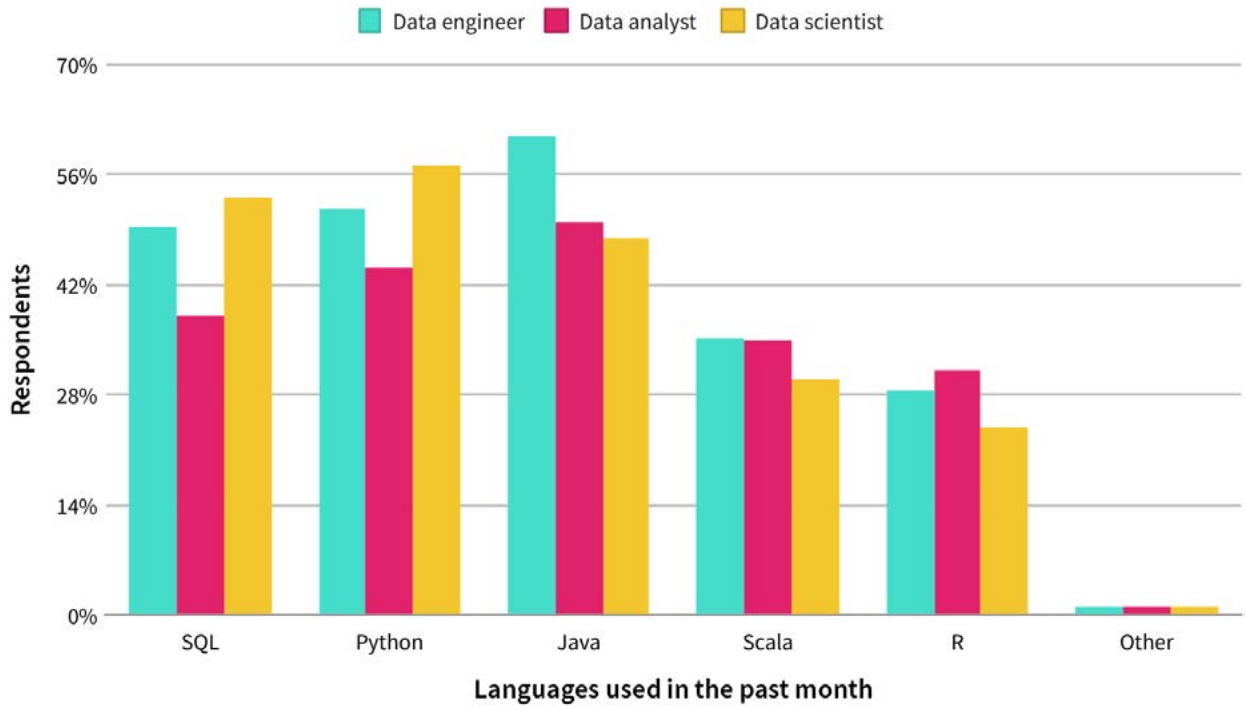
In the aggravate, **Java** was the language of the data team, with recent use by 55% of respondents. **Python** and **SQL** were close behind, in recent use by 52% and 48% of data practitioners, respectively. **Scala** (33%) and **R** (27%) were runners-up.



When broken out by job title, the overall patterns held, although there were some notable changes.

Among data engineers, **Java** was significantly more popular than any other language, with a nine-point lead over the runner-up **Python**, 60% to 51%. Data analysts also favored **Java** but by a smaller margin, with 50% using it in the last month compared to 44% using **Python**.

For data scientists, in contrast, **Python** was most common, with 57% of respondents using it in the last month. The second-most-popular language was **SQL**, used by 53% of respondents. **Java** was third-most-common, used by 48% of data scientists.



This data indicates that a data practitioner can expect to encounter SQL, Python, and Java most commonly, which is unsurprising.

Among data practitioners overall, the spread between these most-common languages was only 6 points, confirming that none of these have become the dominant language of data. Moreover, each of these languages was used by no more than 55% of data practitioners.

Even between data practitioners with similar job titles, then, there is a fairly high chance that the most-often-used language of one individual will be comparatively unused by another. This opens up significant possibilities for a translation gap when trying to communicate practices, configurations, or tests even within the same organization.

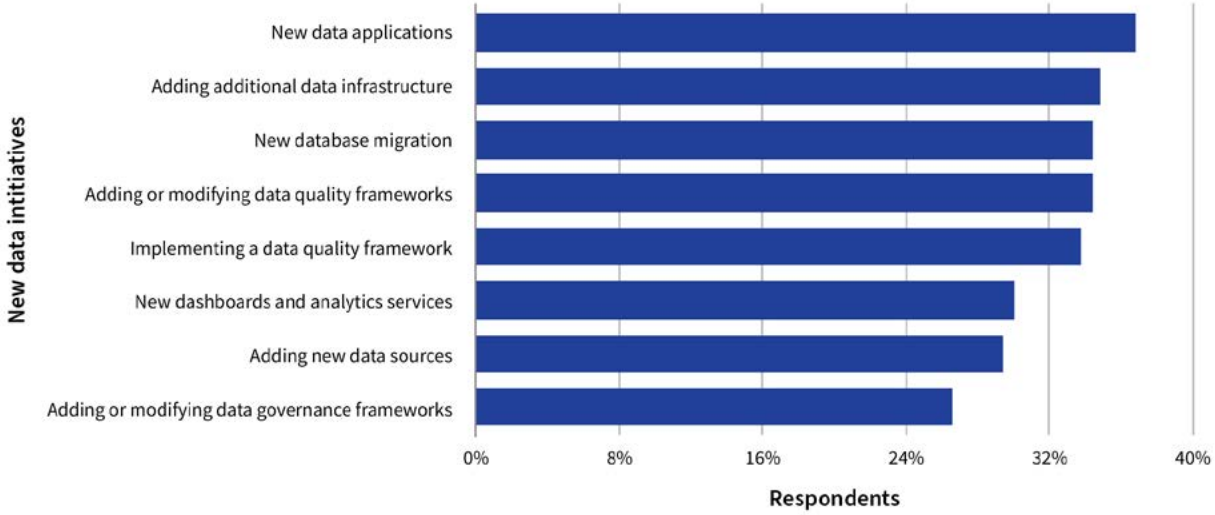
What major data initiatives are you planning?

We asked respondents to select all the major initiatives they were planning around data for the next quarter.

In total, each data practitioner had approximately 2.5 new initiatives planned. The most common was **new data applications**, which were planned by 36% of data practitioners.

New database migrations, adding additional data infrastructure, and **adding or modifying data quality frameworks** were close behind, with 34% of respondents planning on each.

The least common initiative, **adding or modifying data governance frameworks**, was still planned by 26% of data practitioners.



Clearly, the data landscape within organizations is still significantly in flux. This poses a significant challenge to data quality: it's unlikely that any of these initiatives will require no adjustments to a data quality implementation, and each has the potential to require major changes.

Conclusion

Data practitioners are well aware of the need for data quality. Despite strong leadership support, however, actual implementation of data quality lags, with a bare majority of respondents indicating an active data quality program.

We were not surprised to find that a large number of the barriers to data quality fell under the umbrella of communication difficulties. Lack of communication and inter-team understanding were frequently identified as contributors to data quality issues. Data practitioners also called out opaque processes as a major DQ challenge, and there was no single unifying language that was frequently used by a significant majority of data practitioners.

Inconsistencies also play a role, with data practitioners operating on complex technical stacks and having little control over third-party contributors to their data. We do not expect this issue to abate anytime soon, with the average data practitioner having multiple major data initiatives planned for the next quarter.

Finally, data practitioners indicated a less-than-ideal degree of resources available for data quality, with significant proportions lacking the budget and/or prioritization needed to implement data quality; a much higher fraction of respondents are actively carrying out data validation.

Full speed ahead for data practitioners

Despite these challenges, overall data practitioners indicated a high level of trust in their data. Further research is needed to reconcile this trust level with the awareness of data quality issues indicated by other questions. One reason for this disconnect may be the same communication problems and inconsistencies which respondents cited in their descriptions of DQ issues; data practitioners may potentially default to assuming that data they are not familiar with is good.

Overall, our survey indicates that data quality is at an exciting inflection point, with awareness of the issues high but widespread standards, processes, and implementations not yet settled. Data practitioners are seeking a way to alleviate widespread communications barriers and act consistently across variable environments.

Ease of usability and/or acquisition on the part of existing data quality solutions may be a significant issue, as many more data practitioners reported active data validation than reported active data quality as a whole. Data quality offerings that can mimic SaaS data validation services in terms of availability and ease of adoption will fill an important need of the data practitioner community.

About Great Expectations

Great Expectations, our powerful and expressive open source platform for data quality, empowers data practitioners to operationalize their data quality processes while also effectively collaborating with nontechnical data stakeholders. Our mission is to revolutionize the speed and integrity of data collaboration.

Backed by some of the best open source and data infrastructure investors in the industry, GX is building a SaaS product that will augment the GX platform with end-to-end capabilities for creating and automating data contracts.

Get data quality news straight to your inbox.

Sign up for GX's email list to get tips from leading data developers, perspectives on the data quality landscape, news from the GX community, and more.

Sign up today

greatexpectations.io