DATA QUALITY ASSESSMENT: HOW TO GET STARTED

So, you're thinking about doing a data quality assessment. Here's an overview of what you'll need to start and along the way.

For the best results in your data quality assessment, you'll need three main things:

- Persuasive tools
- Technical tools
- A plan

Your goals:

- To figure out your current data quality situation

You'll need:

- Buy-in from higher-ups
- Buy-in from peers and other colleagues whose help you will need
- Budget (possibly)

While it's possible to assess your data quality without any additional money, it'll make things a lot easier. If you find that you need data quality operations like enrichment services, it'll be a requirement later on.

Persuasive Tools

This is an often-overlooked but critical part of your data quality assessment. If your persuasive tools aren't up to scratch, you won't get an accurate idea of your data quality, let alone start to repair it.

The main hurdles you'll be trying to overcome with these tools will probably include:

- Why is this worth your time? (From your boss and higher up)
- Why is this worth my time? (From your peers and colleagues)
- Why should this have any budget?

Here are some examples of persuasive evidence you can use when arguing that it's important to know your organization's data quality situation.

General Statistical Citations

The importance of data quality has been widely recognized, especially in the last few years. What analysts, consultants, or outlets do the people around you trust most?

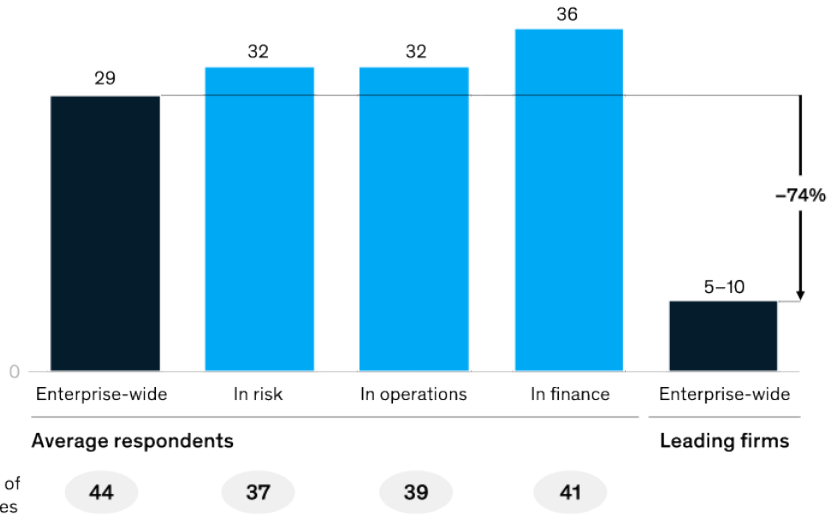
Use those as primary sources. Below are some examples.

From McKinsey

- https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/designing-data-governance-that-delivers-value

Lack of data quality and availability can cause employees to spend a significant amount of time on non-value-added tasks.

Time spent on non-value-added tasks due to poor data quality and availability[1]
Estimated % of total employee time



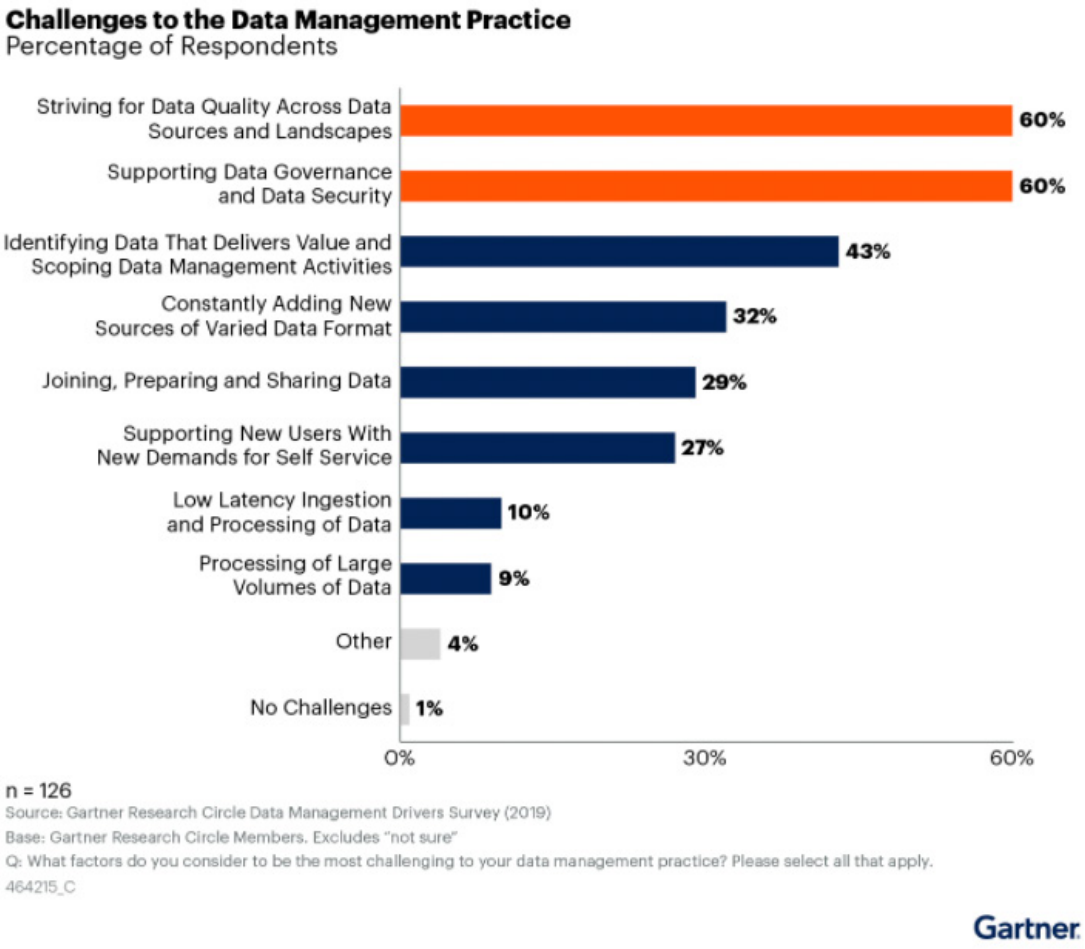
[1] Data sourcing, data aggregation, data reconciliation, data cleansing, manual reporting, etc.
Source: McKinsey Global Data Transformation Survey, 2019

McKinsey & Company

From Gartner

- https://www.gartner.com/en/documents/3986583/cost-optimization-is-crucial-for-modern-data-management-

Data quality is the foundation of everything that is built on an organization's data assets. Poor data quality destroys business value. A recent Gartner survey of reference customers for the forthcoming 2020 edition of "Magic Quadrant for Data Quality Solutions" found that organizations estimate the average cost of poor data quality at $12.8 million per year. This number is likely to rise as business environments become increasingly digitized and complex. Also, another recent survey found that the need to strive for data quality across data sources and landscapes is the joint-biggest challenge to data management practice (see Figure 1 and).



From PriceWaterhouseCoopers, via the Wall Street Journal

- https://www.wsj.com/articles/ai-efforts-at-large-companies-may-be-hindered-by-poor-quality-data-11551741634

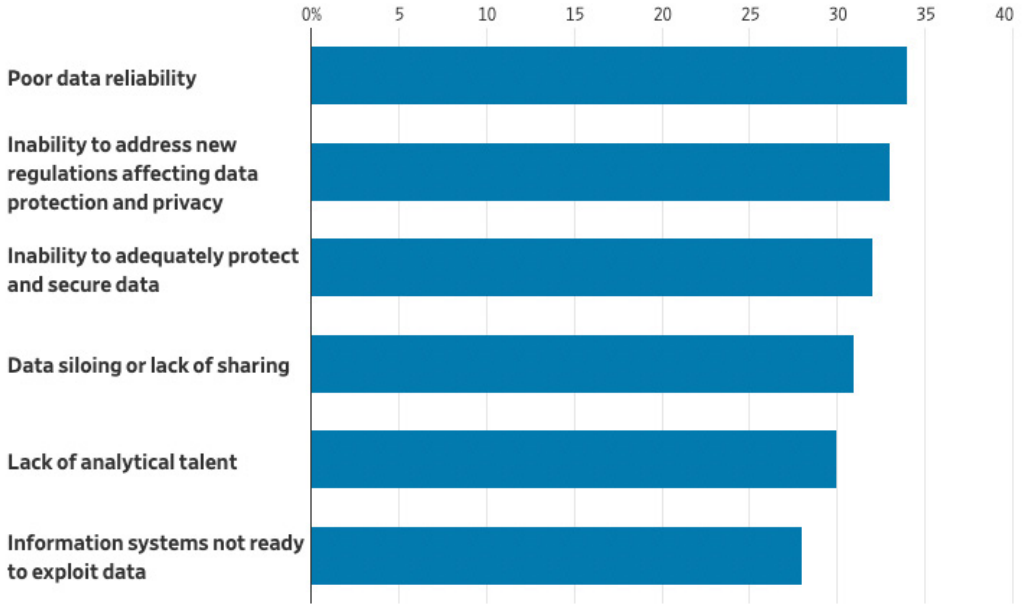
CIO JOURNAL

AI Efforts at Large Companies May Be Hindered by Poor Quality Data

Executives surveyed by PwC said cleaning up their data would lead to big cost savings and revenue gains

Obstacles to Monetizing Data

Executives surveyed by PwC said efforts to extract value from data troves face a number of challenges.



Source: PwC, Trusted data optimization pulse survey, February 2019

Legal Requirements

As data-related regulations—particularly privacy-related ones—become more expansive and more common, they introduce the need for data quality in order to achieve legal compliance.

A data quality assessment can be a good first step for figuring out how much you'll need to add or change to achieve compliance and start forecasting your DQ budget.

Usually regulations don't mandate "data quality" by name: instead, they require a level of knowledge, access, and control over your data that can only be achieved by robust data quality and governance.

Research if there are any of these regulations that apply to you; if so, they'll be a valuable asset to your data quality campaign. Here's a primer and below are some of the most well-known examples.

General Data Protection Regulation (GDPR)

Applies to

Any business, of any size, that 1) has personal data about an EU citizen, or 2) carries out processing of personal data about anyone within the European Economic area

Requires data quality because

Among other things, EU citizens have the right to require you to update, delete, or send them copies of all the data you have about them, on demand.

So not only do you need to be able to find all the data you have about a person, you need to make sure that any changes to it are synced completely across all your systems. You also need to be able to delete it (and prove you did it).

Consequences if you fail

Huge fines:

> *For especially severe violations, listed in Art. 83(5) GDPR, the fine framework can be up to 20 million euros, or in the case of an undertaking, up to 4 % of their total global turnover of the preceding fiscal year, whichever is higher. But even the catalogue of less severe violations in Art. 83(4) GDPR sets forth fines of up to 10 million euros, or, in the case of an undertaking, up to 2% of its entire global turnover of the preceding fiscal year, whichever is higher.*
>
> from https://gdpr-info.eu/issues/fines-penalties/

Resources to start with

GDPR: How is it Different from US Law & Why this Matters? Blog Privacy and Data Security Insight

GDPR fines and data breach survey: January 2021- DLA Piper

California Consumer Privacy Act (CCPA)

Applies to

For-profit businesses meeting at least 1 of the following criteria: 1) Gross annual revenue is greater than $25 million, 2) Deal in the personal information of 50,000+ California residents, households, or devices, and/or 3) Get at least 50% of their annual revenue from selling California residents' personal data.

Requires data quality because

California residents have the right to know what data about them you have, so you must be able to locate it across all your systems. As with GDPR, you also have to be able to delete it upon request. Additionally, you need to have enough control over your data that if a person opts out of having their data sold, you can actually execute that opt-out.

Consequences if you fail

The California Attorney General can pursue civil penalties against you.

Technically, this has been superseded by the California Privacy Rights Act; however, CCPA enforcement is ongoing, while CPRA enforcement will not begin until 2023.

Resources to start with

California Consumer Privacy Act (CCPA) | State of California

The California Consumer Privacy Act: Frequently Asked Questions

California Privacy Rights Act (CPRA)

The CPRA took effect in January 2021, with full enforcement set to begin in January 2023. It will supersede the CCPA.

Requires data quality because

CPRA creates additional consumer rights, with accompanying need for additional technical capacity to support this.

Resources to start with

The California Privacy Rights Act Has Passed: What's in It?

Internal statistics and anecdotes

While some people are really impressed by numbers, others need the real-life context of a story. By having some of both, you can appeal to the broadest possible audience.

Of course, if you can only get one or the other, that's better than nothing.

Here are some examples of statistical information that could indicate a low level of data quality:

- High bounce percentage of marketing outreach (postal, email, text, etc.) and the amount and percentage of the budget associated with those bounced messages: $100k budget with 20% bounce rate = $20k wasted.
- Low customer satisfaction scores, particularly regarding processes under your direct control, like customer service interaction or ease of scheduling.
- Low effectiveness of personalized advertising or targeting compared to generic.
- Low percentage of data routinely used.
- Analytical information that remains steady while business operations change (or vice versa).

Here are some examples of anecdotes that could indicate low data quality:

- Recorded or recalled customer service interactions with customers who are upset about a data quality issue (wrong name, used old address, etc.). This is particularly effective for high-value clients.
- Complaints from data analysts or other users that they do not have access to the data they need.
- Accounts of convoluted internal processes that take up a lot of employee time relative to their benefits.

The core argument is: if poor data quality is manifesting in these ways, it could be having other effects you aren't aware of. A data quality assessment will help you figure out if that's true.

On top of that, having some internal evidence that data quality is already affecting your company is particularly useful if you get a "well, we're different" reaction to either general information or legal requirements–assuming those are even applicable to your organization).

Technical tools

How will you carry out the data quality assessment, if you get the opportunity? Have some ideas lined up!

A critical step is going to be data profiling.

Your data profiling options will vary depending on several factors, including:

- What format your data is in
- How technical you want to be, also known as “how much code are you OK with writing?”
- How much money (if any) you have to spend
- What data you have

Your goal with data profiling is to get a summary-level idea of what your data is. Do you have the data you expect? In the amount you expect? With the values you expect? Following the rules you expect?

No matter what kind of data you have, you will want information like the most and least common values, the frequency distribution of values, most and least common value patterns and frequency distribution of value patterns.

For numerical data, basics include statistical information like the mean, median, minimum, and maximum; for non-numeric data, it’s characteristics like the number of characters in the data, character set(s) used, and character patterns.

On top of this fundamental information, you can also get data about completeness, number of zeros and nulls, business type predictions, form and format identification and statistics, and more.

This is all pretty standard stuff: anything that advertises itself as data profiling should be able to deliver these. It’s factors like “how technical you want to be” that will determine what options you have.

Outcome

The final piece you’ll need is an outcome: what will the tangible results of your data quality assessment be?

In addition to the profiling results, a strong deliverable for your assessment is a plan of attack for improving and/or further investigating your data quality.

This plan isn’t an end-to-end description of everything you’ll do in its entirety. In fact, we strongly recommend against that.

Instead, it should be a short list of the key systems (or tables) you want to focus on first. You should select these candidate options because you believe you will be able to carry out a tightly-scoped data quality project on them in a relatively short time.

This will let you execute the trailblazer approach to data quality, and transition from your initial data quality assessment into beginning to build a strong, ground-up data quality program.

Good luck on your data quality assessment!