**How to Create a Data Quality Framework**

**Step-by-step instructions on identifying key areas to focus on**



Photo by [Clément Dellandrea](https://unsplash.com/@clementdellandrea?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText) on [Unsplash](https://unsplash.com/s/photos/frame?utm_source=unsplash&utm_medium=referral&utm_content=creditCopyText)

What happens when your data arrives later than expected? Or the values in a column that should be populated are all NULL? Data quality issues like this are often lagging indicators that something in your data pipeline went wrong. By the time they occur, the damage has already been done and you…

Here are a few common data quality issues and their potential causes:

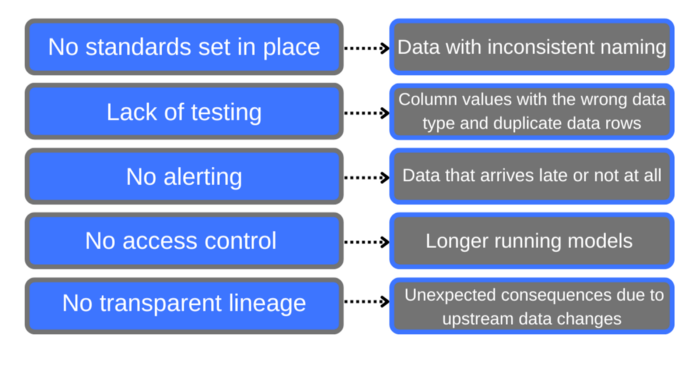


Image by author

Focusing on data quality initiatives is essential for producing data that the business can depend on. Without certain tools and processes in place, your data will be hard to trust. Nobody will know for sure if the data is leading to accurate insights. And, if you can’t trust the data, what can you trust?

Implementing quality checks will also allow you to decrease the time you spend troubleshooting issues that pop up. Instead, you can be proactive and foresee any problems *before* they happen and disrupt workflows. The best way to emphasize data quality is to create a data quality framework for your team to follow.

**What is a data quality framework?**

A data quality framework is a tool for organizations and data teams to define different data quality characteristics so that they can work towards developing systems and processes that help improve these characteristics. This framework focuses on seven key areas of data quality- accuracy, completeness, timeliness, uniqueness, governance, lineage, and consistency.

**Accuracy** refers to how close the actual values of your data are to their true value. Do the characteristics of the actual value make sense compared to the true value? For example, is the credit card number column 16 digits like it’s supposed to be?

**Completeness** looks at whether there are missing values in a data table. Is there a column or row that is missing a value? If it is, is this expected?

**Timeliness** looks at the freshness of your data. How often is new data being ingested? Is data available when your team needs it?

**Uniqueness** checks for duplicate data rows. Are there two rows with the same primary key? How do their other values compare?

**Governance** refers to who has access to your data and how secure it is. Everyone should have the minimum access needed to do their jobs. PII data should be hidden from users. You also want to ensure your data is meeting regulations in your industry.

**Lineage** tracks the movement of your data from ingestion to orchestration. Ideally, you should be able to see dependencies between models and how a source column is being transformed downstream.

Lastly, **consistency** looks at the patterns in your data. Is your data being ingested and transformed at the same time every day? Or is this variable? Is the volume of data being ingested similar every week?

**Why should I create a data quality framework?**

A data quality framework allows you to pinpoint the areas of quality that will have the greatest effect on the business. It follows the idea that “80% of the results come from 20% of the effort”. When you take the time to focus on the areas that are causing the most issues, then you will have the maximum impact on your organization.

A framework gives you a better understanding of the areas you are lacking in. It also helps you gain support around investing in data quality because you will be able to present exactly what you are working on to stakeholders clearly and concisely. A framework allows you to present a problem and a solution that can help prevent that problem from ever occurring again.

**How to create a data quality framework**

So, how do you create the framework on that you will base all of your data quality initiatives? Let’s walk through the steps.

1. **Record all of the known data quality issues you’ve had in the past or suspect to have in the future.**

Think about past incidents that you’ve worked on or issues you’ve run into. For example, if your data pipeline failed 3 weeks ago, this would be an issue that you would want to right down. If data was missing from one of your tables for a certain date, you would also want to record this.

**2. Categorize each one into one of 7 categories- accuracy, completeness, timeliness, uniqueness, governance, lineage, and consistency.**

Now you’ll want to revisit the descriptions I listed earlier of the different data quality pillars. Which one does each of the issues you’ve listed fall under? A pipeline failure would fall under timeliness since the data for that day was late to be delivered. Missing data in a table would fall under completeness.

**3. List out each category and put the corresponding items underneath it.**

You want to make sure these issues are properly organized. This will give you something to present to stakeholders or leaders in your organization if they ask about the data quality initiative.

**4. Count which 3–4 categories have the most amount of items.**

You will want to focus your framework on the categories that have the most amount of items. These are the ones that your organization struggles with the most. 80% of your improvement will come from focusing on these. Once you’ve determined these, I recommend creating a simple graphic to remind yourself of what matters.

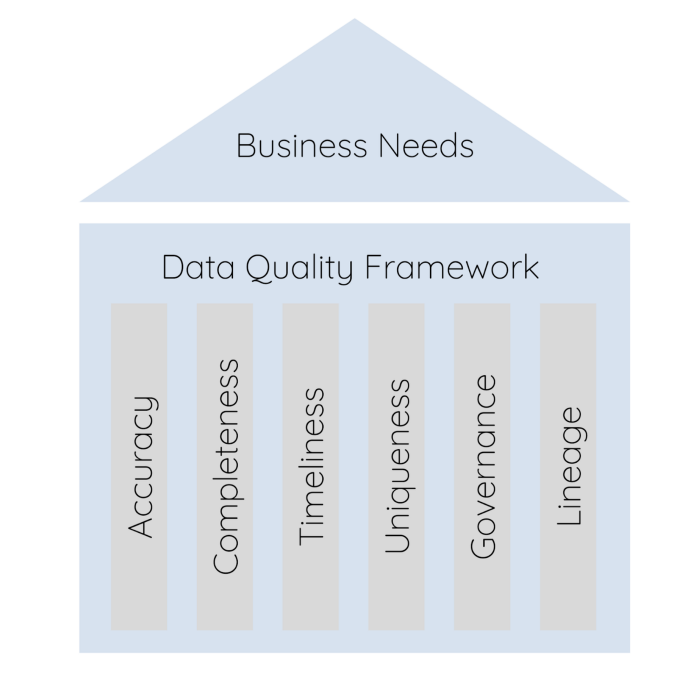


Image by author

**5. List the potential solutions for each of these problems under their corresponding category.**

This will help you to define an overarching goal for each data quality pillar. For example, testing may prove important for increasing completeness and accuracy within your data tables. I recommend brainstorming potential tools like [Datafold](https://www.datafold.com/) and [re\_data](https://www.getre.io/) you want to implement to help with each pillar and thinking about different types of tests like [dbt tests](https://docs.getdbt.com/docs/build/tests), [dbt expectations](https://github.com/calogica/dbt-expectations), or even custom-built ones.

Systems are also just as important as tools and tests when creating solutions. Think about what your team needs to do to make these initiatives successful. Do you need to create an on-call schedule or alerting action plan for when something goes wrong?

**6. Get granular with each of these solutions and list out exact tasks that need to be done for each of these solutions to be successful**

I recommend creating these specific tasks as tickets on [Jira](https://www.atlassian.com/software/jira) so that they can be added directly to a sprint. You want them to be detailed enough that you can point them and immediately start implementing them.

**7. Share the basic framework with stakeholders so they understand what you are working on and why.**

Lastly, share what you’ve done! You are doing this to produce high-quality data that everyone can depend on. Make sure you draw attention to this! Frame the initiative in a way stakeholders will get excited about. For example, adding tests to our data models will alert us when something is broken, allowing us to fix your favorite dashboard *before* it breaks.

**Recommended Tools**

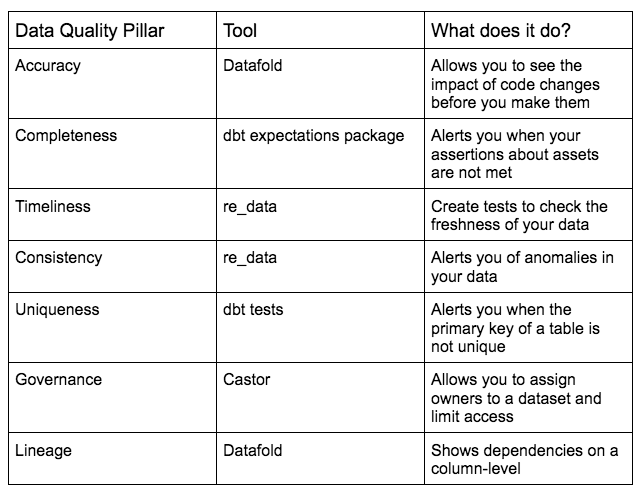


Image by author

**Conclusion**

When creating a data quality framework, you want to focus on accuracy, completeness, timeliness, consistency, uniqueness, governance, and lineage. A framework with these pillars in mind will help to keep your processes organized and ensure you are working on the tasks that deliver the best results.

Luckily, there are a lot of amazing tools that are making it easier to produce high-quality data. The biggest hurdle is often figuring out what you need to do and which of the tasks will have the biggest impact on the business. It’s never too late to implement this type of framework, even if you are in the middle of your data quality work.

Regroup, reorganize, and watch your data quality soar to new heights!

For more on producing high-quality data, check out my free weekly [newsletter](https://madisonmae.substack.com/) on analytics engineering.