3 BEST PRACTICES FOR AUTOMATING VALIDATION OF DATA



When it comes to automating data validation, companies already know the pros and cons.

Before cloud computing ushered in the 21st-century world of big data, data entry was manual and time-consuming, with three major issues: data sets were often too small to be reliable or impactful, human error could easily introduce bad data, and the data may not measure what you intend to measure.

Manual validation of manually-entered data, then, is less a process than a random spot-check and, thanks to human error, your IT team will waste time fixing bad data after they've resulted in bad situations. (Sure, manual data validation can work for small datasets or if you need a lot of specific, hands-on control – but these are rare instances in the life of big data.)

Data validation, when automated, stops bad data from corrupting your data warehouse before it can even get in. More important is that automating data validation actually allows you to work with truly large data sets. The bottom line – there's no reason not to automate your data validation processes. Let's take a look at three best practices to follow when beginning automation.

What is data validation?

When data is collected, it is beneficial, if not downright necessary, to make sure that data has been checked to ensure the quality of that data. If data is bad, business units will hesitate to make decisions around it, questioning whether to trust the data. IT will hesitate to spend time and money

improving data resources. The company at large will suffer, too. Once bad data is in the business stream, it can be used to support decisions that ultimately go awry or communicate poorly with customers. Bad data is estimated to cost companies over $700 billion a year. Per company, that averages to nearly 30% of your revenue.

The goal of data validation is to mitigate these issues via processes that make sure collected data is both correct and useful.

There are plenty of methods and ways to validate data, such as employing validation rules and constraints, establishing routines and workflows, and checking and reviewing data. For this article, we are looking at holistic best practices to adapt when automating, regardless of your specific methods used.

Automating data validation: Best practices

Without further ado, here three best practices to consider when automating the validation of data within your company.

1. Create a culture where everyone values data quality.

Data is not a function or silo of IT. Instead, data is an IT tool that supports any business need.

This philosophy is important when automating data validation: everyone should have a stake in clean, trustworthy data. Adapting a company culture that values the importance of data means every employee has a responsibility for improving data processes, including automation. If a small set of data is found to be poor or incorrect, the IT team shouldn't be blamed. Instead, the situation should be looked at holistically – what data is collected? What business need does the data support? Is it necessary or beneficial? How can we use IT to correct the issue in order to support the business need?

A good rule of thumb when starting any data quality effort is to make sure that effort directly supports a business goal.

2. Ensure your data structure is stable.

Timing is important when automating the validation of data. If your company is developing an aggressive approach to take control of your data, you may be moving too quickly, particularly if your data systems and infrastructure aren't set. Or, if you're a start-up, you may not have a full understanding of exactly the data you need and whether you're collecting it appropriately.

For instance, perhaps your data is currently housed in in various on-site servers as well as across the multi-cloud. Automating validation of this data in its current status could lead to significant vulnerabilities because your infrastructure isn't stable. So, before automating, make sure your data and data warehouse are accurate and useful for the business needs.

Some ways to ensure stability are to maintain single databases when possible (the more transferring or referencing outside of a database, the more unexpected errors may occur) and to isolate your data validation to reduce risk of contamination.

3. Introduce a data steward.

Just because you've automated data doesn't mean you never need to check in on the process. On the contrary, data automation means that your IT team can better spend time improving processes and mitigating issues before they have a major impact on the business. A data steward can be essential to this process, owning responsibility for data validation.

A data steward is someone who ensures defined data processes are running smoothly, tests data quality when automated alarms go off, and develop front- and back-end checks that are easy. Plus, there's no need to reinvent the wheel. Some tried-and-true methods include:

- **Using real-time data.** The entire point of big data is to use the most up-to-the-second data, so outdated and bad data is gone. Real-time data guarantees that everyone is operating under the same information, and this prevents a few people making a decision based on a tiny or inaccurate amount of data.
- **Collecting statistics.** Use your data to inform your data. Tracking statistics means you can set alarms based on trends or patterns. For instance, if one day's load volume is suddenly half its normal rate, an automated alarm can alert the data steward to investigate.
- **Validating both input and target sources.** Your developers can build automated ways to develop assessment control that reviews new data before it gets added to the warehouse. For instance, if a user is trying to enter a birthdate into an age field, the infrastructure will know the input is incorrect. With target source validation, devs can incorporate rules that compare end data to the source data to ensure accuracy, especially when data transformation necessarily occurs.

Keeping these best practices in mind when automating your data validation can ensure a smoother transition with less down time and contribute to an overall appreciate for quality data within your company.